THE [illegible] OF VISCOUS AND [illegible] OF THE [illegible]

[illegible] Aerospace Laboratories

This report has been published by the Army Information Office, [illegible] and is available to the National Technical Information Service (NTIS). At NTIS it will be available to the general public, including foreign nations.

This report has been reviewed and approved for publication.

APPROVED: [signature]

UNCLASSIFIED

SECURITY CLASSIFICATION OF THIS PAGE (When Data Entered)

REPORT DOCUMENTATION PAGE	READ INSTRUCTIONS BEFORE COMPLETING FORM
1. REPORT NUMBER: RADC-TR-76-399	2. GOVT ACCESSION NO. / 3. RECIPIENT'S CATALOG NUMBER
4. TITLE (and Subtitle): THE EVALUATION AND SYSTEMS ANALYSIS OF THE SYSTRAN MACHINE TRANSLATION SYSTEM.	5. TYPE OF REPORT & PERIOD COVERED: Final Technical Report. Nov 74 — Aug 76
	6. PERFORMING ORG. REPORT NUMBER: N/A
7. AUTHOR(s): T. C. Halliday, E. A. Briss	8. CONTRACT OR GRANT NUMBER(s): F30602-75-C-0078 new
9. PERFORMING ORGANIZATION NAME AND ADDRESS: Battele Columbus Laboratories, 505 King Ave, Columbus OH 43201	10. PROGRAM ELEMENT, PROJECT, TASK AREA & WORK UNIT NUMBERS: 31025F, IDHS0427
11. CONTROLLING OFFICE NAME AND ADDRESS: Rome Air Development Center (IRDT), Griffiss AFB NY 13441	12. REPORT DATE: January 1977
	13. NUMBER OF PAGES: 82
14. MONITORING AGENCY NAME & ADDRESS (if different from Controlling Office): Same	15. SECURITY CLASS. (of this report): UNCLASSIFIED
	15a. DECLASSIFICATION/DOWNGRADING SCHEDULE: N/A

16. DISTRIBUTION STATEMENT (of this Report)

Approved for public release; distribution unlimited.

17. DISTRIBUTION STATEMENT (of the abstract entered in Block 20, if different from Report)

Same

18. SUPPLEMENTARY NOTES

RADC Project Engineer: Nicholas M. DiFondi (IRDT)

19. KEY WORDS (Continue on reverse side if necessary and identify by block number)

Machine Translation Evaluation
Stem Dictionary Update
Semantic Expression Dictionary Update

20. ABSTRACT (Continue on reverse side if necessary and identify by block number)

This report is the product of contractual effort to determine how machine translations produced by SYSTRAN can be improved upon and where to integrate the required correctional procedures. Two tasks were pursued: the first a survey of SOTA in translation evaluation methods; the second, the actual determination of translation deficiencies and identification of correctional procedures. The survey did not produce any new methods of evaluation, and it was already known that the old procedures tested quality characteristics but did not

CONT.

DD FORM 1 JAN 73 1473 EDITION OF 1 NOV 65 IS OBSOLETE

UNCLASSIFIED 407080

SECURITY CLASSIFICATION OF THIS PAGE (When Data Entered)

identify deficiencies. Consequently, a method was adopted which did meet the requirements to achieve the objectives of the contractual effort. Russian source material from Mechanical Engineering and Metallurgy were selected for translation using a base-line SYSTRAN configuration. One sample of each was translated and then corrected by a bilingual expert in each field. Two types of corrections were considered implementable, stem dictionary update and semantic expression update. The same samples were re-translated incorporating first the updated stem dictionary and then both it and the updated semantic expressions dictionary. Improvements to sentences under each condition were recorded. Also, a different sample from each field was translated by the base-line configuration and each updated version to maintain control of carry-over effects that may occur due to updating. Overall results across both subject areas show a 50% improvement in the sample of sentences when the stem dictionary is updated and 56% when the semantic expression update is included in translation. Carry-over effects indicate 40% improvement vs 41% improvement respectively can be achieved in translating related material.

END

REPORT SUMMARY

The research described in brief in this summary and in detail in the final report entitled "The Evaluation and Systems Analysis of the SYSTRAN Machine Translation System" was conducted under Contract No. F30602-75-C-0078 for Rome Air Development Center. This research was directed at the evaluation of a machine translation system, SYSTRAN; SYSTRAN is used by the Technical Translation Group of the Foreign Technology Division of Air Force Systems Command for the rapid translation of scientific and technical literature from the source language, Russian, to the target language, English. The purpose of the research was to examine existing, off-the-shelf machine translation evaluation methods for applicability to raw SYSTRAN output, select the one thought most suited for that purpose, evaluate raw SYSTRAN output on the basis of the selected method, and use the results of the evaluation for a systems analysis study, particularly in the area of postediting processes.

The search for existing, off-the-shelf machine translation evaluation methods was done by extensive literature searches, both of U.S. and of foreign literature, and by mail contact with U.S. and foreign institutions presently active in machine translation research. The evaluation methods thus found were analyzed by a team of researchers in terms of applicability to SYSTRAN and probable obtainable results. All of them were rejected for various reasons, including impracticability, previous questionable results, and prior use with SYSTRAN. In their place an entirely new approach to the problem of machine translation evaluation was proposed. (Cf. Final Report, Section II, "Review of Existing Machine Translation Evaluation Methods"; Appendix A, Bibliography).

The new approach involved the evaluation of raw SYSTRAN output based on the improvement achieved through system dictionary updating. Raw translations of technical literature were produced on an operational SYSTRAN system. These translations were edited by bilingual subject-matter experts and those emendations suited to lexicographic updating were incorporated into the system dictionary of an experimental SYSTRAN system. Additional translations were produced on the experimental system and were compared with the translations obtained from the original system. At the same time parallel translations of related, previously unedited technical material were produced to study the carry-over effects of the changes in the system dictionaries. Computer programs were used for the comparison of two related translations. The computer programs produced printouts of sentences differing in the two translations for human analysis and evaluation. (Cf. Final Report, Appendix B, "SYSDIF Computer Program Description"). Significant results were obtained in the improvement of raw translation output of the technical fields under study in this research. Improvement was as high as 56 percent of the total number of sentences in the translation samples. This improvement carried over to a lesser, but still significant, degree in the translations of the related technical

fields. Of the number of sentences in this second sample, over 40 percent was improved. At the same time, a cost-effectiveness study of this process demonstrated the value of concentrated lexical editing by bilingual subject-matter experts for immediate and for long-range benefits to a machine translation system. (Cf. Final Report, Section III, "Design and Application of an Experimental Approach to Machine Translation Improvement and Evaluation"; Appendix C, "An Experimental Measure of the Cost Effectiveness of Expanding SYSTRAN Subject Area Dictionaries").

Several additional conclusions could be drawn from the results of the current research, which dealt specifically with the translation of two technical areas, metallurgy and mechanical engineering. It can be assumed that an iterative process incorporating the above methodology of updating system dictionaries on the basis of raw system output would rapidly improve a given technical lexicon to its practical limit. It can further be assumed that the same type of research would yield similar improvement in technical areas not covered in the present research. Furthermore, although not all emendations made by the bilingual subject-matter experts could be used for system dictionary updating, the remainder of these emendations was analyzed and produced suggestions for further improvement in several other components of the SYSTRAN system. These components included the set of lexical routines which are an extension of the system dictionary, and several of the syntactic and semantic routines of the system. Other areas affected by this analysis were the input preparation process and the various postediting processes presently in use with the SYSTRAN machine translation system. (Cf. Final Report, Section IV, "Conclusions and Recommendations").

TABLE OF CONTENTS

		Page
I.	Introduction	1
II.	Review of Existing Machine Translation Evaluation Methods	3
	Introduction	3
	Classification of Machine Translation Evaluation Methods	4
	Preliminary Feasibility Analysis of Existing Evaluation Methods	17
	Applicability Study of Existing Evaluation Methods	19
	Review of Suggested Alternate Approaches to Machine Translation Evaluation Methods	22
III.	Design and Application of an Experimental Approach to Machine Translation Improvement and Evaluation	25
	Introduction	25
	Description of Proposed Method	26
	Preliminary Data Analysis	29
	Comparative Data Analysis	32
IV.	Conclusions and Recommendations	42
	Introduction	42
	Discussion of Problem Areas	42
	Discussion of Evaluation Methodology	47
	Conclusions and Recommendations for the Lexical Component of the SYSTRAN Machine Translation System	49
	Conclusions and Recommendations for the Integrated SYSTRAN Machine Translation System	52
Appendix A.	Bibliography	59
Appendix B.	SYSDIF Computer Program Description.	61
Appendix C.	An Experimental Measure of the Cost Effectiveness of Expanding SYSTRAN Subject Area Dictionaries	63

EVALUATION

The procedures used to correct deficiencies in translations produced by SYSTRAN; namely, stem dictionary update and semantic expression dictionary update were already known to be methods of improving the quality of the output. The value of the work performed under this contract is that it provides quantitative degrees of improvement. Since the level of lexical capability was known for each of the two subject areas addressed in this effort, the Air Force can extrapolate these results to similar situations and in conjunction with the cost data provided can make more accurate assessments of the cost benefits to be derived in implementing these procedures. This work is in support of the written word exploitation mission as defined in TPO No III A.

Nicholas M. DiFondi

NICHOLAS M. DIFONDI
Project Engineer

SECTION I

INTRODUCTION

This research report describes a 20-month effort to evaluate and to improve the translation quality of the SYSTRAN system for machine translation of Russian scientific and technical literature. The evaluation of the translation quality was based on the unedited output of the SYSTRAN system; research toward the improvement of translation quality was directed primarily at improvement of the existing SYSTRAN dictionaries, in the linguistic sense. No attempt was made to restructure this lexical component of the system. Similarly, no attempt was made to analyze or to restructure the syntacto-semantic components of the system. Results of this research were two-fold: (1) the thorough analysis and subsequent updating of the lexical component resulted in a marked improvement in the technical accuracy of the translation quality in the unedited output; (2) a number of constructive ideas was generated for further possible improvement of SYSTRAN, both in the lexical and in the syntacto-semantic components of the system.

The SYSTRAN research effort was specifically directed toward a number of well-defined objectives. The primary goal was to explore the possibility of establishing an evaluation method for machine translation output based on objective criteria. Such a method was to be chosen from existing machine translation evaluation methods. In addition to simply evaluating unedited machine translation output, the selected evaluation method was also to be used for systematic feasibility studies for further improvements to the SYSTRAN system, and for objective evaluation of the results of experimental changes introduced into the various components of the system, as well as studies of carry-over effects of limited changes from one component to other parts of the system. The selected evaluation method was also to be applied to systems analysis studies of SYSTRAN, both in a language-oriented and in a computer-oriented sense. These studies included comparisons of unedited and postedited output for the linguistic part, and studies of trade-offs between linguistic improvement and computer usage for the computer-oriented part.

The specific objectives of the SYSTRAN research effort were accomplished in two major tasks. The first of these tasks was an exhaustive study of existing machine translation (MT) evaluation methods. This study involved a preliminary analysis of such evaluation methods for their possible applicability to the specific requirements of SYSTRAN. This first task also involved a comparison of those MT evaluation methods that were found to be applicable to SYSTRAN with the method presently in actual use for the evaluation of SYSTRAN output. Based on these analyses, an evaluation method was selected for experimental use in the remainder of the research effort. The second major task was the employment of the selected evaluation method in the analysis of SYSTRAN, mainly in its lexical component, but also, to a certain degree, in the related systems level. This analytic work concentrated mainly on the benefits achievable in translation quality

through systematic efforts in dictionary updating, with some tentative feasibility studies on the usefulness and desirability of such major concentrated systems updating.

SECTION II

REVIEW OF EXISTING MACHINE TRANSLATION EVALUATION METHODS

Introduction

Although mechanical translation or machine-aided translation of natural language texts had its inception practically with the introduction of the electronic computer, much less research has gone into the development and study of relatively objective machine translation evaluation methods. The search for existing, off-the-shelf MT evaluation methods was conducted in two concurrent phases. The first of these phases was an extensive and exhaustive search of pertinent literature, through information dissemination facilities such as ERIC and NTIS. This literature search included review of technical magazines such as those published by the Association of Computational Linguistics, and of books by preeminent authors in this technical field, such as H. P. Edmundson and Paul Garvin. However, literature of other technical areas was also reviewed. These areas included artificial intelligence, simulation, computer-aided instruction, modern language teaching, information processing, and experimental psychology. In addition, study of pertinent Russian, French, and German literature in natural language applications also formed part of this preliminary research, which covered the period of the mid-1950's to the present.

The second phase of this search was direct contact with institutions presently active in the field of machine-aided translation, both in the United States, such as the Universities of Texas and California, and outside the United States, such as the University of Montreal in Canada and the University of Grenoble in France. Disappointingly, it was found that published research of MT evaluation methods had ceased in the mid- to late-1960's, and, although a surprising amount of research was to be found in machine-aided translation, none of it seemed to be directed toward evaluation methods for the results of this research.

Nevertheless, the search for existing MT evaluation methods led to approximately thirty types of methods. Included in this number were both theoretical, untested methods, and methods that had actually been applied to the evaluation of machine-aided natural language translations, with varying degrees of success. Also included in this number were several methods of evaluating language capacity not directly related to MT evaluation; they were included because it was felt that they could, with only slight modifications, be applied to MT evaluation work.

The initial analysis of existing MT evaluation methods showed that, in general, they tended to evaluate a translation by measuring a single component of the translation on some suitable scale. For a more thorough analysis of their applicability and usefulness as regards the

SYSTRAN system, these evaluation methods were classified into related groups based on the evaluated components, such as readability, comprehensibility, intelligibility, and informativeness. The following outline is a description of MT evaluation methods arranged in precisely that form. It is listed here partly for general information. Further details can be found in the references listed in the bibliography in Appendix A. It is also listed here, however, because this initial analysis formed the basis of the more thorough analysis by Battelle researchers who were involved in the effort to single out those evaluation methods thought to satisfy best both the criteria established for a satisfactory MT evaluation method and the remaining requirements of the overall research effort.

Classification of Machine Translation Evaluation Methods

The eight major headings in this classification combine MT evaluation methods that form logically homogeneous groups. These groups contain both untested theoretical methods and experimentally tested ones. It is to be noted that Heading VII and Heading VIII contain major subordinate and possibly applicable methods, respectively. In addition, although many of these evaluation methods are theoretically applicable to general natural language work, the following discussion will be restricted, whenever possible, to the context of the translation of restricted technical areas.

I. Subjective Evaluation Methods

In general, a subjective evaluation reflects the judgment of a bilingual expert, based mainly on his knowledge of the two languages in question. However, personal preferences may occasionally introduce a slight bias in their conclusions.

A. Free Subjective Evaluation

1. Fidelity of Translation

This method is a completely subjective evaluation of a translation. It is done by a bilingual expert who compares the translation to the original source material and who bases his evaluation on the correctness of the information transferred from source to target language. This method is listed as a primary reference, since it, in effect, is the method presently used to evaluate SYSTRAN output.

2. Ouality of Translation

This evaluation method is practically the same as the one described under fidelity of translation. However, in addition to a consideration of correct information transfer, the evaluator also considers such other factors as use of correct syntax and language style. Because of these additional factors, it is not so easily applicable to MT evaluation as the simpler fidelity of translation test. The arbitrariness of completely subjective evaluation can be reduced either by using a team of evaluators to obtain an average consensus of opinion, or by restricting the evaluator within specific limits, as in the following section.

B. Scale-Restricted Subjective Evaluation

1. Fidelity of Translation

a. The first evaluation method under this heading, discussed by Miller and Beebe-Center, is a subjective evaluation of the information content of a machine translation. It can be done by comparing the mechanical translation with either the original material or a criterion translation of the original. The evaluator, however, is asked to assign a grade of 0 to 100 to the translated material, theoretically forcing him to a more careful examination of the translated material.

b. A similar method, discussed by Crook and Bishop, uses the same theoretical procedures, but reduces the scale of evaluation to a restricted range of 1-25. An adaptation of these concepts has been used for SYSTRAN evaluation, in which, in addition to the free subjective evaluation discussed above, a very general 5-point scale has been used by SYSTRAN evaluators.

2. Quality of Translation

a. Crook and Bishop, using the same scale of 1-25 applied to fidelity of translation judgments, also conducted experiments on the evaluation of several types of translations, mechanical and human, without reference either to a

criterion translation, or to the original source material. The scores in these experiments showed strong correlation with the fidelity of translation tests. Once again, it is obviously possible to reduce the subjective aspect of these evaluations by using teams of evaluators in addition to point scales.

II. Lexico-Syntactic Evaluation Methods

The methods discussed here use lexical and syntactic units to evaluate translations. It must be stressed, however, that these units are used not as concepts with specific functions within a grammar, but as independently manipulable units.

A. Lexical Evaluation

1. This method, discussed by Miller and Beebe-Center, is based on lexical units. Criterion and test translations are compared on the basis of a common vocabulary, i.e., the number of same words appearing in the two translations. The evaluation score, N, is defined as follows:

$$N = \frac{S}{T}$$

where S is the total number of identical words in the two translations, and T is the total number of words in the criterion translation. Limited experiments showed reasonable correlation of evaluation scores with acceptable translation quality.

2. A second evaluation method employing lexical units, also described by Miller and Beebe-Center, again is based on the number of shared words in a criterion and a test translation. The evaluation score depends on the ordinal position of the shared words in the respective texts, and is defined as the ratio of the number of shared words in the same ordinal position to the number of shared words. Limited experiments were conducted to test this method; none of these was on machine-aided translation research.

B. Syntactic Evaluation

1. The same theoretical concept that was used in the first of the lexical evaluation methods is applied by Miller and Beebe-Center to develop another possible evaluation method. The comparison of criterion and test translations is based on a preselected number and type of syntactic constructions, such as an adjective-noun combination. The evaluation score, N, is defined as above in terms of shared and total number of the preselected syntactic units. Results from experiments on this method were described both as tentative and as inconclusive.

2. The following method, also based on the research of Miller and Beebe-Center, would employ the linguistic concept of degree of grammaticality for evaluation purposes. It is suggested that a rating scheme based on this concept could be developed and used with source material and test and criterion translations. However, the authors state it only as a theoretical possibility. It is not developed to the testing stage.

3. A final syntactic evaluation method by the same authors is based on another linguistic concept, that of an immediate constituent analysis of both source material and target translation. The evaluation would be based on a comparison of the two analyses. This method is also suggested only as a theoretical possibility, and is not developed to the testing stage.

III. Evaluation Methods Based on Measures of Transmitted Information

One important factor in the evaluation of translations of natural languages is the amount and the correctness of the information transferred from the source to the target language. The evaluation methods discussed under this heading all reflect emphasis on that factor.

A. Evaluation Methods Based on Information Theory

1. Miller and Beebe-Center discuss the possibility of an evaluation method based on the concept of reversibility. This method involves a translation from a source to a target language, and a translation of this material back to the source language.

The evaluation would be based on a comparison of the information content of the original source material and the translated translation. This method, too, is not developed beyond the theoretical discussion.

2. The same authors suggest the application of information theoretic concepts due to C. Shannon to natural language evaluation. In the proper notation, the method is based on the following formula:

$$T = H(x) - H_y(x), \text{ where}$$

T = total information common to two translations, x and y

H(x) = amount of information in x

$H_y(x)$ = amount of information in x when y is known.

Using Shannon's method of calculating H, H(x) is obtained by letting a test subject guess successive letters in translation x; $H_y(x)$ is obtained by letting a test subject guess successive letters in x, after having read translation y. Now, if x and y are translations, T, a function of H, should be high when translations are good, and low when they are bad. Some tentative experiments on natural language translations showed precisely this correlation. It must be pointed out that this method, using, as it does, alphabetic units, really measures information transfer from an information theoretic and not from a linguistic sense. The following section lists evaluation methods that are more language-oriented.

B. Evaluation Methods Based in Informativeness Scales

1. This evaluation method uses the concept of information transfer and is based on an informativeness scale developed by Carroll. Test subjects study first a test translation and then either the original source material or a criterion translation. They then assign an evaluation rating to the source material or the criterion translation based on how informative they considered it to be in comparison with the test translation. The underlying assumption is that, if the translation is good, the original material will rank low on the informative-

ness scale, and it will rank high if the translation is bad. Carroll's scale used in this research is a ten-point scale, with "9" defined as "extremely informative" at one end, and "0" defined as "the original contains, if anything, less information than the translation" at the other extreme. It is to be noted that the concept underlying Carroll's scale and its application is precisely Shannon's information transfer adapted to natural language research. In this particular experiment, Carroll's scale was applied to individual source and target language sentences and showed the expected correlation.

2. The same method, i.e., Carroll's scale and the necessary source and target language test material, was used in experiments conducted by Leavitt, Gaber, and Shannon, with one difference. These researchers felt that simple sentences gave insufficient context for judging informativeness. They therefore used textual units rather than individual sentences as their test material. A textual unit is defined as a block of text treating one commplete idea or concept. These researchers also found the expected correlation between their test material and the ratings obtained from Carroll's scale.

3. Although based on a somewhat different concept, the evaluation method developed by Szanser can also be discussed under the heading of information transfer. Szanser developed a scale based on the concept of usefulness, i.e., it is a scale developed with emphasis on a monolingual user of a mechanical translation and its degree of usefulness to him. Evaluation of a translation is made in two steps. First, test subjects evaluate machine translations completely subjectively and without reference to source material. Second, test administrators assign usefulness scale ratings based on their interpretation of these subjective evaluations. The usefulness scale itself is a nine-point scale, with "8" defined as "fully adequate" and "2" defined as "poor". Odd scale points, like 7, are defined in terms of "between 6 and 8", and the logical end points 10 and 0, "absolutely perfect" and "absolutely no sense", are not used. This scale was used to evaluate MT work in England, and, in general, the user-oriented scale rated MT output as "average".

IV. Evaluation Methods Based on Measures of Intelligibility

Two factors used to evaluate mechanical translations are intelligibility and comprehensibility. In general usage, these terms are considered synonymous. For the purpose of this analysis, these terms will be differentiated. In the context of MT evaluations, intelligibility will refer to the ease with which a mechanical translation can be understood, i.e., how clear is it to a reader?

A. Evaluations Based on Sentences

1. Carroll, in developing the informativeness scale for MT evaluation, also developed a similar scale for the measure of intelligibility. Test subjects considered individual, out-of-context sentences drawn from translated material only, and rated these sentences for intelligibility using Carroll's nine-point scale, where "9" is defined as "perfectly clear" and "1" is defined as "hopelessly unintelligible". In several of these experiments a time measure was included in connection with the intelligibility evaluation. This time measure showed almost linear negative correlation with intelligibility ratings.

2. An alternate evaluation method, also based on this concept of intelligibility, consists of a sequence of tests called "judgment of clarity" tests by their author. Developed and tested by Pfafflin, this method requires test subjects to evaluate sets of single, out-of-context sentences on a very general three-point scale, with the points defined as "clear, unclear, no meaning". However, the term "unclear" includes both sentences that are difficult to understand because of a poor translation and those that are unclear because of an ambiguous construction. The sequence of tests consists of an application of this rating scale to several sets of sentences. One set consists of hand-translated sentences, the second set consists of machine-translated ones, and the final set is a mixed set drawn from the hand-translated and the machine-translated sets. One interesting result of these experiments was that in the mixed set of sentences the hand-translated ones were consistently ranked higher, and the machine-translated sentences were consistently ranked lower on the clarity scale than in their unmixed sets.

B. Evaluations Based on Units Other Than Sentences

1. Leavitt, Gates and Shannon, in addition to applying Carroll's scale of informativeness to MT evaluation work, also conducted experiments with Carroll's scale of intelligibility. Again, the unit used for testing purposes was not a single sentence but a textual unit. The results, however, showed no appreciable variations from those derived by Carroll himself, in his work.

2. Another team of researchers, Crook and Bishop, basing their work on the same factor of intelligibility, developed their own scale for its evaluation. The Crook and Bishop intelligibility scale consists of seven points, rather than nine; "1" is defined as "about as good as comparable material in the target language" and "7" is defined as "only vague impression of meaning can be obtained". In this case, test subjects were asked to evaluate complete texts, rather than single sentences or textual units. Therefore, in this test, context played an important part in determining intelligibility. This conclusion is further strengthened by an additional result. The intelligibility tests were administered in conjunction with comprehension tests, and showed generally more correlation with comprehension than with fidelity of translation measures.

V. Evaluation Methods Based on Measures of Comprehensibility

Methods of MT evaluation based on the factor of comprehensibility reflect one of the more important properties of a translation. Where the factor of intelligibility, discussed above, is based on the general clarity of a translation, whether in its entirety or in out-of-context segments, comprehensibility is based on how thoroughly an entire translation can be understood.

A. Comprehension Based on Direct Questions

1. The most obvious method for testing comprehension is to ask questions about the material to be comprehended. Miller and Beebe-Center suggest this approach as a possible testing method for the comprehension of mechanical translations. No tests were run on this method.

B. Multiple Choice Comprehension Tests

1. A second evaluation method for comprehension is the well-known multiple choice test. As in the case of the direct question tests, Miller and Beebe-Center suggest the multiple choice test as a possible evaluation method for mechanical translation. However, except for suggesting some possible guidelines for developing and administering such tests, the authors do not develop this method further.

2. Another researcher in the area of mechanical translation evaluation methods does apply the multiple choice method. Pfafflin developed and administered such a method, which was applied as follows. The multiple choice tests were prepared based on the original source material. Test subjects were tested both over hand-translated and over machine-translated target language material. Results showed that, although scores for both types of translations were above the guessing level, differences in correct responses between the hand and the machine translations were significant. So were the differences in time spent on each. It must be carefully noted that one important result in this research was that both scores and times improved on machine translation tests as test subjects became accustomed to the peculiarities of machine-translated material. In addition to taking the multiple choice tests, subjects were asked to grade the adequacy of machine translations. The scale used was a three-point scale divided into "adequate; adequate as a guide for deciding whether to request a better translation; useless". Like Szanser's approach, this scale is directed toward the user of a translation, and, like the average results obtained by Szanser, Pfafflin found that fully 86 percent of the machine-translated material fell into the "adequate as a guide" category.

3. The same type of multiple choice tests was used by Crook and Bishop in MT evaluation studies. Great care was taken in the selection of test subjects. They included experts on technical fields, as well as nonexperts; they included proficient bilinguals, as well as monolinguals. It was found that differences in scores among the several types of hand and machine translations were more significant

than other factors like knowledge of a technical field or proficiency in both source and target languages.

4. Still another researcher, Orr, also used multiple choice tests in his MT evaluation work. His work is of interest because of the additional controls he used. Orr based his multiple choice tests on hand-translated, rather than source language, material. However, Orr initially analyzed his multiple choice tests, generating supporting statistics like item difficulties, item test correlations, and Kuder-Richardson reliabilities. Orr also based his evaluation scores on three different types of multiple choice questions. One test consisted of direct or literal questions, based on material explicitly stated in the texts; a second test was composed of equivalent questions, based on material covered in the text, but paraphrased; a final test consisted of indirect or inferential questions, based on material not covered explicitly in the text, and consequently requiring an understanding of the text beyond a single word or sentence. In general, test scores varied for the different test types to a statistically significant degree.

5. It is important to note that the evaluation method most often applied to MT research, multiple choice tests, was not developed specifically for MT evaluation, and thus the statement reflects the state of the art for MT evaluation work. The multiple choice evaluation method is varied slightly by researchers through the introduction of additional, supplemental tests. Thus, for example, the material developed by Orr was also used by Leavitt, Gates, and Shannon in their work of evaluating machine translations. The only additional improvement over the work by Orr was the testing of the reading ability of test subjects to form as homogeneous a sample of test subjects as possible. The test used for this purpose was the Nelson-Denny Reading Test. Research results, however, to a large degree paralleled the results of other researchers using the same basic evaluations method.

VI. Evaluation Methods Based on Readability

Like the MT evaluation methods based on the factor of comprehensibility, evaluation methods based on the concept of readability must consider, if not the whole, then at least a sizable segment of the translated material. This requirement is due to the fact that the method, although called a readability method, measures the appropriate overall contextual cohesiveness.

A. The Cloze Technique

1. The evaluation method measuring readability was first applied by Crook and Bishop to their research on machine translations. The method itself, called the Cloze technique and developed initially by Taylor, is based on the psychological concept of Gestalt, leading to the consideration of a mechanical translation as a theoretical, unified whole. In its linguistic application the method is used in the following manner. The text of a given translation is mutilated by the omission of a certain number of words, and test subjects are asked to fill in the blanks. The number of correct responses determines both the score and an evaluation based on that score. Several variations can be introduced into this method. Words can either be omitted on a random basis, or words can be omitted every nth step, where n can be an arbitrarily chosen value. Scores can be kept based on responses of the exact original omitted word, or synonyms and paraphrases of the original can be permitted. Crook and Bishop used elimination of every 8th word. They also ran two tests. In one, correct responses were only those reproducing the original word; in the other, anything that reproduced the sense of the original was accepted. Tentative results showed some correlation between quality of translation and readability. Reading time, measured as an independent variable during the administration of the Cloze technique, also showed the same correlation.

2. Two other researchers, Sinaiko and Klare, also applied the Cloze technique in an extended series of tests to evaluate mechanical translations by the readability factor. Abstracts of papers published by the two authors indicate that readability tests based on this technique have been

successful to a certain degree, but that better results can be achieved by combinations of methods. In at least one sequence of evaluation experiments, Sinaiko and Klare used this technique in conjunction with reading comprehension tests and clarity ratings, as well as time needed to complete each of these tasks.

VII. Miscellaneous Subordinate Evaluation Methods

In general, the evaluation methods listed under this heading were all used in conjunction with other methods and to supplement information gained from these other methods. They are listed here separately for the sake of logical clarity and logical completeness.

A. Time-Related Evaluation Methods

1. In the study of machine translation evaluation using multiple choice tests conducted by Orr, he measured two additional variables, rate of work, and rate of accuracy. Rate of work was defined as the total time taken for a given multiple choice test in terms of the number of 10-minute periods into which the test was divided. Rate of accuracy, on the other hand, was defined as the number of items correctly answered per 10-minute period. These two additional measures were used together with the results of the multiple choice tests for a comprehensive evaluation of machine translations.

2. Researchers for Arthur D. Little, Inc., published the results of an evaluation of a machine translation system and the method applied used reading time as the major independent factor. Comprehension of the test material was considered only secondarily. The surprising and questionable conclusion drawn from this test was that there were no appreciable differences between hand-translated and machine-translated material.

B. Other Dependent Methods

Leavitt, Gates, and Shannon added a factor called a Task Importance Rating to their study involving intelligibility and informativeness tests for MT evaluation. This rating has two sections. The first section is to be completed by test subjects only if the test subjects recall ever having seen or used the information in the textual unit being evaluated, or if the test subjects decide that they could have used that information in their areas of technical

competence. The first section of the Task Importance Rating consists of seven questions, each to be graded on a nine-point scale on the relative importance of the translated material to the test subjects. The second section of the Task Importance Rating, which must be completed, contains three questions, each of them again to be graded on a nine-point scale on the relative usefulness of the textual unit in presenting factual information, or in helping to find related technical material. This rating was used in conjunction with scores from the intelligibility and the informativeness scales for the evaluation of machine-translation material. In general, however, it can be concluded that methods of this type may contribute to the over-all evaluation of MT material, but that they are not particularly suited to be applied independently for that purpose.

VIII. Possible Alternate MT Evaluation Methods

The following section includes descriptions of tests or references to methods that were not developed for application to machine-translation evaluation. They also were not used for this purpose. In general, not enough information was obtained on these methods for an applicability study of these methods of MT evaluation. Nevertheless, they are here listed as eventual possible additional evaluation methods, thus completing this outline of theoretical, practical, and possible MT evaluation methods.

A. Psychologically-Oriented Tests

1. A psychologically-oriented language proficiency test based on a technique known as Clozentropy has been developed by Darnell. This test measures the performance of a test subject in terms of a group norm. For example, foreign students learning English have been rated according to the extent to which their responses to the test questions agree with the normal or average responses to native speakers. Here a good possibility exists of rating machine translations by equating the machine translations as the norm of native speakers, and scoring the machine translation according to the method of the Clozentropy technique.

2. Another language evaluation technique, also psychologically-oriented, is one developed by Spolsky, who used the principles of a noise test to evaluate the language proficiency of foreign students.

The method consists of an oral test of understanding of English sentences to which noise levels of 1 db, 4 db, 7 db, 10 db, and 50 db are added electronically. The greater the electronic interference, the better a language knowledge must be for good understanding of the distorted sentence. For the purpose of MT evaluation, degrees of understanding of a set of distorted machine translations, at a constant db level, could be tested for correlation with quality of translations.

B. Statistically-Oriented Tests

1. Nakamura has developed a procedure for the automatic identification of natural languages on the basis of small samples of text. The procedure is based on a statistical treatment of language elements and has been applied to approximately 25 languages from a variety of language families with reasonably good success. Therefore, this procedure could perhaps be used to differentiate between good and bad translations. That is, if the procedure could be made sensitive enough to classify a submitted sample as "definitely English, possibly English, no decision possible", then it could be used for the purpose of MT evaluation work.

This last section completes the review of existing, off-the-shelf machine translation evaluation methods. The next section will consider these methods, group by group, with a view to their applicability for evaluating a specific machine translation system, SYSTRAN.

Preliminary Feasibility Analysis of Existing Evaluation Methods

The existing machine translation evaluation methods, described briefly in outline form in the previous section, were subjected to a thorough and comprehensive feasibility study by a team of researchers consisting of a mathematical linguist, an expert bilingual editor, a behavorial psychologist, and a statistician. The feasibility study centered on an analysis of the general properties of the evaluation methods and their possible applicability to the specific requirements of the present MT research project. This analysis was constrained by the following general considerations.

Theoretically, a scientific procedure should be objective, valid, and reliable. A practical method for evaluating translations of

natural languages with any degree of confidence, should satisfy, to as large an extent as possible, those three conditions. It should be valid, i.e., it should measure whatever it is designed to measure with an acceptable degree of accuracy. It should be reliable, i.e., it should produce the same results with an acceptable degree of consistency. It should be objective, i.e., it should be as free of subjective bias as possible. However, because this particular procedure deals with the evaluation of natural language output, it cannot be completely objective, for the final judgment on translation quality must still be made by a human bilingual expert. Nevertheless, the procedure should be able to offer the human evaluator as much objective information about a translation as possible. A secondary consideration, not so important as the three main considerations, should be ease of application, if a procedure is to be used repeatedly.

The preliminary analysis of existing MT evaluation methods brought out two important points. The first point was that most of the existing evaluation methods tend to isolate one component of a translation, such as comprehensibility, and to evaluate the translation on the basis of that single component. The second point was that, in general, there exists only one method for measuring such a component. Other related methods tend to use the same theoretical concept, with variations either in the basic definition of that component, or with differences in the units used to measure the component, coupled with differences in the interpretation of the resulting evaluation scores. This second point does have, however, an important result: the fundamental concept underlying an evaluation method based on one component has been used often enough in experimental studies, so that it can be assumed to be both reasonably valid and reliable.

However, the Battelle research team felt that human evaluators and, more importantly in the context of this research, users of MT material judge the quality of a translation by more than just one preselected component. Therefore, the research team involved in this preliminary analysis of evaluation methods based its selective work on two factors. The first of these was a close adherence to the criteria of validity, reliability, and objectivity discussed above. The second of these was a careful study of the possible mutual interrelationship or logical interaction of those evaluation methods selected at this stage of the analysis. The analysis was performed on groups of evaluation methods sharing an underlying theoretical concept; the decisions and the reasons for the decisions following this analysis are discussed below. These results are presented in outline form for each group of analyzed evaluation methods. The groups themselves are based on the groups defined for the initial description of these methods, and follow the order of the outline in the preceding section.

Applicability Study on Existing Evaluation Methods

Direct consideration of the off-the-shelf machine translation evaluation methods included the facts that these methods included both theoretical, untested methods, and experimentally tested ones, as well as both questionable and acceptable results from this latter group. Of the existing methods based on single components, those selected were chosen both on the basis of the relative importance of that component for translation evaluation, and on the basis of a strong possibility of close interaction among these single components.

I. Subjective Evaluation Methods

Evaluation methods based on subjective evaluation were eliminated immediately. Two main reasons can be cited for their exclusion. The first is the aim of as much objectivity as possible, lacking in this approach. The second is the fact that this method is presently used to evaluat SYSTRAN output, and part of this research was to find, if possible an alternative evaluation tool.

II. Lexico-Syntactic Evaluation Methods

Evaluation methods based on lexico-syntactic principles, as defined in the previous section, were also immediately eliminated from further consideration. There are a number of reasons for their exclusion. Some of these methods are theoretical in nature and have never been tested. Others have been tested under such limited conditions as to make their reliability and their validity questionable. Most important, however, is the fact that many of these methods, such as the one based on ordinal word positions, are linguistically unmotivated, even though positivie correlation may exist between experimental scores and translation quality.

III. Evaluation Methods Based on Measures of Transmitted Information

The evaluation method based on concepts due to Shannon measures information transfer as defined in communication theory, not as in natural language communication. It is also a very cumbersome, impractical process. However, the same concept, linguistically oriented, underlies the methods using informativeness scales. Of the ones discussed, the method developed by Carroll is probably the most valid and reliable because of repeated use. Correct information transfer is one of the most important properties of a translation; informativeness was therefore the first of four factors to

be chosen for possible MT evaluation work. It was also felt that single sentences out of context formed samples too small for evaluation pruposes. The Carroll scale of informativeness was therefore selected to be applied to textual units.

IV. Evaluation Methods Based on Measures of Intelligibility

Intelligibility of a translation, as defined for purpose of MT evaluation, i.e., ease of understanding, was selected as a second factor for judging translation quality. Of the three intelligibility scales discussed in conjunction with evaluation methods, Carroll's is again the most reliable and valid, due both to repeated use and the careful effort with which it was constructed. It will therefore be used to measure the component of intelligibility. However, as in the case of informativeness and for the same reasons, the scale will be used in conjunction with a textual unit rather than with single sentences.

V. Evaluation Methods Based on Measures of Comprehensibility

Comprehensibility, again as defined for MT evaluation, i.e., thoroughness of understanding, was felt to be another important property of a translation, and was therefore selected as the third of four factors to be considered. In existing methods this component has been measured only by multiple choice tests. Hence, for reasons of validity, applications of this method must use multiple choice tests. For thorough testing, it was felt that both the literal type question and the indirect type question should be used, in order to eliminate guessing levels, at least in part.

VI. Evaluation Methods Based on Readability

It was decided to include a fourth factor in the study of MT evaluation, that of readability. This factor was included mainly for a theoretical study of possible effects on interaction among the major components selected. One testing technique only has been applied to the component of readability – the Cloze technique. It was to be applied to the testing of readability under the following conditions: regular interval elimination of words from a sample test, and acceptance of paraphrases of the original text.

VII. Miscellaneous Subordinate Evaluation Methods

The minor MT evaluation methods under this heading were all eliminated from additional consideration, except for the time variable. Although the importance of time as an independent variable may be questionable, it is a relatively simple matter to gather the appropriate data. It was therefore decided to

measure time as a factor, with its possible inclusion in the final analysis of MT evaluation methods to depend on later studies of its correlation with the other selected components of a mechanical translation.

VIII. Possible Alternative Evaluation Methods

The evaluation methods discussed under this heading were initially included as possibly applicable to MT evaluation work. Since none of them had been applied to actual MT evaluation, and since therefore no data were available on factors like reliability and validity, none of these methods was seriously considered for further use in the present project.

From the above study it was concluded that, given the state of the art of MT evaluation techniques, a combination of methods testing for the four factors of informativeness, intelligibility, comprehensibility, and readability would probably yield the most reliable tool for valid MT evaluation. These methods would be administered singly, using the directions suggested by their developers. The results, obtained from several technical fields and several types of translations, would then be evaluated several ways, including statistical tests of significance. The same results would then also be analyzed as a group for interaction and dependence trends, using automatic interaction detector processes, also primarily statistical in nature. In these preliminary stages, one additional minor suggestion was discussed. It was felt that both the Cloze technique in the readability tests and the multiple choice questions in the comprehensibility tests could reasonably be supplanted by appropriately designed, easily applied scales such as the ones used in the informativeness and intelligibility tests developed by Carroll. These two alternate approaches to the present MT evaluation problem – direct use of existing tests or slight modifications of some of these tests for the sake of convenience – were then extensively studied for advantages and disadvantages and the resultant finding, with adjoint recommendations, were then submitted to the Technical Translation Group of the Foreign Technology Division in Dayton for approval. The following general considerations were also raised in conjunction with this analysis.

No translation, whether human or machine-aided, will ever be considered completely satisfactory by human evaluators, although the translation of technical material can approach that goal more closely than translation of, say, classical literature. It follows, therefore, that, with or without objective evaluation criteria, given the present general state of MT quality, human translations will be better than post-edited machine translation, which in turn will be judged better than unedited machine translation. Consequently, the question of an evaluation method

for machine translations should perhaps be limited to evaluating the output of an MT system based on the design criteria and limitation of that system, rather than on a theoretically perfect linguistic output. Finally, in conjunction with this last point, consideration must also be given to the goal of a translation system. In the case of a system like SYSTRAN whose function is the rapid translation of Soviet technical literature, the main criterion for evaluation should be the correct transfer of information content of the source material to the target language, in a form most practical to the potential user of that system.

Review of Suggested Alternate Approaches to Machine Translation Evaluation Methods

Two alternate approaches to machine translation evaluation, based on existing methods, were developed. The first included the use of tests for factors like intelligibility, both individually and in groups, for such evaluations. These tests were to be used as originally developed by their authors. The second approach included the same factors; the difference from the first approach was due to minor modifications in scoring some of these factors. These two suggested alternate approaches to the evaluation of a specific MT system, SYSTRAN, can be studied joinly in terms of their advantages and disadvantages.

One immediate advantage of these approaches is the fact that the selected tests had all been used effectively in previous MT evaluation work, hence, they could on that basis be considered valid. Since, in general, they had produced positive results, i.e., the results of these tests had correlated well with projected results repeatedly, they could also be considered reliable. These methods, therefore, satisfied two of the desired criteria for a scientific process, reliability and validity. In addition, the use of strictly limited scales and objective grading of multiple choice tests introduced a certain amount of objectivity, thus satisfying, at least in part, the third required criterion for a scientific process. There were, of course, other possible advantages. The use of an automatic interaction detector could determine dependencies among the main components to be considered, thus possibly reducing the number of components, and thereby simplifying the overall method. Finally, although the number of test subjects used to obtain results from the various tests would have to be large enough to be statistically significant, in general, the test subjects could be monolinguals, rather than expert bilinguals.

Balanced against these favorable considerations were those that could be considered possible disadvantages. Since all of these evaluation tests had been used before – several of them, in fact, with the SYSTRAN system – the strong possibility existed that another application of these methods would yield no new results, even if the several methods were considered as one group. Furthermore, automatic interaction detection techniques obviously do not guarantee to produce dependencies among test factors, thus not necessarily resulting in a reduction of the number of

components. Neither can such a technique guarantee to select, if dependencies exist, the most practical or the most convenient method in terms of time and test material requirements, or in terms of number and type of test subjects needed. It is such requirements that make the use of several of the selected evaluation methods impractical for repeated application. Yet for repeated use of MT evaluation, ease and convenience of use must be considered important factors of an evaluation method. Finally, there is a major weakness that practically all of the discussed evaluation methods share. Although results obtained from these methods may correlate well with quality of translation, many of them do not really test the correctness of translation, the basic purpose both of an MT system and of an evaluation method for that system.

Because of these considerations but with special concern for the last point raised, a further alternate approach to MT evaluation was proposed by Battelle, based on the initial research condition: that if no existing, off-the-shelf MT evaluation method were found to be acceptable, the method presently in use at the Technical Translation Group in Dayton would be used for the actual systems analysis. However, the suggested method did involve some adaptation of this existing method.

Battelle researchers felt that a thorough analysis of the method in use at the Technical Translation Group, a subjective evaluation of MT material, usually by a single expert bilingual, could lead to an adaptation of this method in order to introduce a certain degree of objectivity. The suggestion was to develop a point scale, similar to the ones developed by Carroll for his informativeness and intelligibility tests. The scale would be based on established psychological principles, and would be constructed after a thorough review of the editing processes in use at Dayton coupled with in-depth interviews of MT editors at the Technical Translation Group. This analysis would form the basis for the proposed evaluation scale.

As in the study of the two initially suggested evaluation methods, a certain number of advantages and disadvantages presented themselves immediately. The main advantage of this last method would be the fact that it would test the correctness of a given translation, since the proposed scale would take into consideration both the source and the target language material. The method would be relatively easy to use, since it was to be based on a method familiar to the Technical Translation Group analysts. The method could easily be adapted to indicate either the need or the desirability of various degrees of post-editing procedures in use at Dayton. The introduction of a strictly defined scale would superimpose a degree of objectivity to the existing subjective evaluation. Since the subjective evaluation over the period of its use had been accepted as both valid and reliable, the newer adaptation of this evaluation method could, at least in theory, also be assumed to be valid and reliable. Hence, this newer method also satisfied the requirements of a scientific procedure.

However, these requirements also formed the major drawback of the suggested method. In order to ensure that the method did, in fact, satisfy these requirements, a certain number of tests would have to be run to establish the required correlation between the old method and its adaptation. In addition, the weaknesses of the old method would obviously carry over to its adaptation. These would include the relatively time-consuming effort of a careful evaluation of the material being tested, and the requirement that an evaluator be an expert bilingual.

All of these approaches to MT evaluation – the origina grouping of the four selected main components, informativeness, intelligibility, comprehensibility, and readability, into one group; the minor variation of this original grouping with the introduction of scales for the comprehensibility and readability tests; and the alternate approach based on the modification of the present method in use at the Technical Translation Group in Dayton – were considered in depth with respect to SYSTRAN evaluation. The final decision reached was that, in terms of time and cost requirements and the goals of this research, the concomitant efforts to implement any of these methods would not justify the expected results. Included in the consideration of the total effort were tasks involving the production of required test material, administration of tests to carefully selected groups of test subjects, and the extended analysis of the final results.

As a result of these considerations, an entirely different approach was suggested by the Technical Translation Group in Dayton. This method was designed to accomplish several objectives simultaneously. It would, first, improve components of the present SYSTRAN system; it would indicate the degree of possible improvement to related components of the system; it would yield an evaluation of the system based on this degree of possible improvement. The description of this method, its application to SYSTRAN evaluation work, the obtained results, and the conclusions drawn from the results, all form the second major part of this research report.

SECTION III

DESIGN AND APPLICATION OF AN EXPERIMENTAL APPROACH TO MACHINE TRANSLATION IMPROVEMENT AND EVALUATION

Introduction

The new approach to machine translation evaluation suggested by the Technical Translation Group of the Foreign Technology Division in Dayton was based mainly on existing techniques and systems capabilities in use at Dayton, both in the language area and in systems work. In the language area it involved an extension of a process for improving translation quality through lexical updating necessitated by raw translation editing. This extension consisted in analyzing, at one time, a large corpus of raw data, more completely than is usually done in the post-editing processes. That is, the editing was done with none of the restrictions usually imposed by SYSTRAN design parameters. In the computer systems area, existing programs were used to create experimental SYSTRAN systems incorporating the results of the extensive editing of the raw data, for comparative testing purposes.

In general terms, this new approach was based on an iterative evaluation scheme. The operational SYSTRAN system was used to produce test translations of carefully selected source language technical material. The unedited translations were analyzed for errors, and all editing thought necessary was done. These editing suggestions were analyzed for possible inclusion into existing system stem or semantic expression dictionaries. Two experimental SYSTRAN systems were generated, one including the updated stem dictionary, the other including both updated dictionaries. The same source material was retranslated by the two experimental systems and the raw translations compared against the original translation for qualitative and quantitative raw translation improvement.

This approach involved the careful consideration of two additional factors of importance to this kind of language improvement evaluation. The first was the choice of source language test material to be analyzed for maximum benefits achievable through lexical updating. The second was the equally important choice of test subjects to do the actual editing of this material. The consideration of these two factors involved close cooperation between personnel from Dayton and Columbus. The entire process is described in detail in the following section.

Description of Proposed Method

The SYSTRAN machine translation system is designed to translate source language material from approximately twelve technical areas ranging from mathematics to biology/medicine. Based on past experience with SYSTRAN, technical personnel of the Technical Translation Group in Dayton has rated these technical areas in terms of translation quality of the translations produced by SYSTRAN. The rating scale is a five-point scale with the following markings: good, above average, average, below average, and poor. It was felt that for the experimental purpose of this research those technical areas should be selected that had in the past offered the most problems to the machine translation system, in terms of the wide scope of the technical field itself and the degree of jargonization present in the technical language. Two technical areas were selected. One was mechanical engineering, the only area rated "below average"; the second was metallurgy/metals working, selected from a group of technical areas all rated "average". Because of the scope of this technical area, it requires one of the most extensive vocabularies in the SYSTRAN system. In addition, the sample from this field was limited to the specific area of metals working; mechanical engineering, however, was effectively covered in its entirety.

Two samples for test purposes were selected from each of the technical areas. The first was used for the actual editing and upgrading of the systems dictionaries. The second sample from each area was used for a study of carry-over effects of these upgraded dictionaries. In any analytic work with natural language one problem is to obtain, first, relatively representative samples, and second, if samples are to be compared or contrasted, to match the type and difficulty-grade of linguistic problems in the two samples. In the current research, the samples were made large enough to be considered representative and the material for initial and carry-over effects studies was picked from the same technical journals for as much uniformity as possible. Each experimental sample contained approximately 50,000 words, resulting in a test corpus of 200,000 words. The material for the area of metals working was selected from the Russian journal Kuznechno-Shtampovochnoye Proizvodstvo; the material for the mechanical engineering field came from the journal Vestnik Mashinostroyeniya.

The next important factor to be considered was the choice of editors to evaluate the translation of the source language test material. In theory, to achieve a good translation, a translator should translate into his native language. It was decided, however, that since this research involved evaluation of an existing translation, and since the aim was to achieve the highest degree of technical accuracy possible,

the editors should be expert bilinguals, specialists in the technical fields, and native speakers of Russian. The two men selected for this purpose possessed these qualifications. Both of them are native speakers of Russian; one holds a degree in Mechanical Engineering from the Belorussian State Polytechnic Institute, Minsk. The other holds a degree from the College of Metallurgical and Mining Engineering in Czechoslovakia. Of equal importance, however, is the fact that both of these men have extensive experience in natural language translation. Their combined experience includes work for the Library of Congress, Voice of America, and McGraw-Hill Publishing Company.

These two men were responsible for the editing and the evaluation of the initial test samples of 50,000 words each in the two technical areas. In order to eliminate any psychological bias in evaluating the carry-over effects of the editorial analysis done by these two editors, a group of five researchers was selected for this purpose. This group consisted of native speakers of English, all with degrees in the Russian language, and all with some experience in translation work. The coordination of both parts of this evaluation effort was directed by a theoretical linguist with experience in computational linguistics.

SYSTRAN is not a static translation system. It is updated on a continuous basis, as data are gathered through the various post-editing processes of raw translations. The two initial translations of the source language test samples were produced on the SYSTRAN translation system in actual operation at the inception of the research project. A copy of that operational system was produced, to be held constant through the period of the research project, for reference purposes, if necessary.

The initial raw translations in mechanical engineering and metals working were given to the two editors for analysis and editing. Two approaches can be taken in the editing of a translation produced by a machine translation system. The first is to perform the editing within the parameters of the system itself, disregarding all those emendations that could not be incorporated subsequently into the system. This approach obviously requires a thorough knowledge of the machine translation system itself. The second approach is to consider the raw translation as a natural language segment with no restrictions, make all the emendations thought necessary or desirable, then analyze these corrections, and incorporate into the system all those changes that can be handled by the system. The second approach was used during this research. The editing process itself was a careful, thorough comparison of the raw translation against the original Russian language text.

Upon the completion of the editing process, the set of emendations was analyzed by technical personnel from the Technical Translation Group in Dayton and all changes considered valid were incorporated into the experimental SYSTRAN system. "Valid", in this context, was defined as "possible to be implemented in the lexical component of the SYSTRAN machine translation system". Those changes accepted for inclusion into the system dictionaries were further analyzed for inclusion into either the stem dictionary or the semantic expressions dictionary. An entry in the stem dictionary is one single source language word with its corresponding translation in the target language; a semantic expression is a group of two or more words in which a meaning change occurs because of the relationship of the particular words to each other. Two additional experimental SYSTRAN systems were generated, one including the updated stem dictionary only, the second including both the updated stem dictionary and the updated semantic expression dictionary. The initial source language test samples were retranslated by the two additional experimental systems. At the same time the material for the carry-over effects study was translated using the three experimental SYSTRAN systems. This material was identified as follows: Phase I, initial translations of both the original and the carry-over effects study material; Phase II, the same material translated by the SYSTRAN system containing the updated stem dictionary; Phase III, the same material translated by the SYSTRAN system containing both the updated stem and the updated semantic expressions dictionaries.

To facilitate the evaluation of these additional translations, a computer program was developed to match sentences in any two translations and to print out those sentences that differed in the two translations. There exist four logical categories in such a comparison.

Category One: No changes were made by the editors in a sentence of the initial translation, and no changes occurred in the corresponding sentence of a retranslation. This category includes all sentences that were judged satisfactory in the initial translation.

Category Two: No changes were made by the editors in a sentence of the initial translation, but some change occurred in the corresponding sentence of a retranslation. This category defines carry-over effects in the initial test samples.

Category Three: Changes were made by the editors in a sentence of the initial translation, and the changes appeared in the corresponding sentence of the retranslation. This category includes all sentences with successful updates.

Category Four: Changes were made by the editors in a sentence of the initial translation, but no changes appeared in the corresponding sentence of a retranslation. This category includes those sentences for which the suggested emendations either were not incorporated into the system or were not properly processed by the system.

The computer program developed to compare sentences was able to print out those sentences of categories two and three, but did not identify these by categories. It did not handle categories one and four, even for identification purposes. Additional information on the computer program can be found in Appendix B of this report. This computer program was used to generate comparisons of the following material: initial translations (Phase I material) against Phase II material, for both the initial test material and for the carry-over effects study; initial translations against Phase III material, again for both sets of source language samples. The computer comparisons of the original test sample translations were returned to the bilingual subject-matter experts for their evaluation. The corresponding material for the carry-over effects study was evaluated separately by the group of five Russian-language translators.

The data resulting from the evaluation of the several types of translation were then analyzed to determine the effects of each type of lexical updating on the translation quality of the output of a machine translation system.

Preliminary Data Analysis

The SYSTRAN machine translation system translates from source to target language on a sentence-by-sentence basis, with no reference to preceding or succeeding material. In effect, the system generates computer records that are complete sentences, ready for printout. The processing computer program, developed for evaluation purposes, compares these records in corresponding translations, and prints out

for human evaluation those sentence pairs that differ in the two translations submitted for evaluation. The evaluation of these sentence pairs was done on the basis of a single sentence context, with reference to the source language original when necessary. The actual evaluation procedure will be described in detail, and the results of the analysis of the evaluation will be discussed for the combined corpus of data, and for each of the technical areas separately. A similar procedure will be followed for the material evaluated under the carry-over effects study, and a detailed study of the results of a comparative analysis of the results of the initial material and of the carry-over effects material will also be presented. There are, however, some preliminary results that need to be discussed before the actual evaluation analysis.

These preliminary results concern the raw data gathered during the actual initial editing process; the raw data, a corpus of 100,000 words was evenly divided between metals working and mechanical engineering. The bilingual subject-matter experts who were assigned to the respective technical areas submitted a total of approximately 6,400 corrections. Although not all of these emendations were later used to update the experimental SYSTRAN system, it is nevertheless informative to look at the resulting gross error rate. Since not all of the suggested corrections fall into the lexicographic area, this error rate cannot be calculated on the basis of the word count of each test sample. Two ways exist for looking at this rate. One is based on the operation of the system itself. It generates records that are complete sentences and analyzes these for translation purposes. The other is based on the viewpoint of the potential user who must use the computer output for his purposes. The following tabulations are based both on units of pages and of sentences. The error rate based on sentences is given for the total test sample of 100,000 words, since the breakdown into the two test areas was not available.

ERROR RATE PER COMPUTER PAGE

	Number of Errors	Number of Pages	Errors per Page
Metals Working	3,909	714	5.4
Mechanical Engineering	2,551	602	4.2
Total	6,460	1,316	4.9

The total test sample resulted in approximately 7,000 sentences. The error rate per sentence is as follows:

ERROR RATE PER SENTENCE

	Number of Errors	Number of Sentences	Errors per Sentence
Total	6,460	7,060	0.9

These tabulations immediately reflect one important fact, and that is a difference, possibly only apparent or accidental, from the evaluation of the two technical areas made by the Technical Translation Group of the Foreign Technology Division. Metals working translations were rated "average"; mechanical engineering translations were rated "below average". It would be expected that the calculated error rate would reflect this rating. However, the error rate, based on errors per page, reverses this rating: 5.4 errors/page for metals working, but only 4.2 errors/page for mechanical engineering. Although the text samples were large enough to minimize possible variations in text difficulty, it is still possible to list at least three factors that could have influenced this result. The first is, of course, the possibility that, in spite of the sample size, the quality of the source language material for either or for both technical areas differed enough from the expected norm to account for this result. The second is the expertise in editing of the subject-matter text evaluators and their individual interpretations of the types of corrections to be made in the test samples. In connection with this second factor it must be remembered that the sample for mechanical engineering covered the entire technical field, while the sample for metallurgy was restricted to metals working. Finally, it must be assumed that the ratings assigned to the various technical fields, based on translation quality, are mean ratings, with wide variations possible within each technical field.

Nevertheless, although these error rates are informative, they are also misleading, and must be interpreted carefully. The error rate per sentence, 0.9, is an average and does not mean that almost all sentences in the test samples were in need of correction. Furthermore, not all of the 6,400 corrections were lexical in nature. When these proposed corrections were analyzed, approximately 2,100 corrections were eliminated as being beyond the scope of lexical updating. However, the lexical component of a translation is one part of a machine translation system that affects a potential user of a machine-aided translation, both in terms of inaccurate or erroneous translation of source language terms, and in terms of incomplete lexica that do not

translate terms at all. The following tabulation, based on lexical errors only reflects, first, the completeness of the lexica for the two technical areas, and the possible value of a translation for a user. In this tabulation the error rates are based both on page count and on word count, since the table reflects lexical data. The number of errors includes both inaccurate or erroneous lexical entries and unfound lexical entries.

LEXICAL ERROR RATE

Number of Errors	Number of Units	Error Rate
4,300	1,316 pages	3.3 error/page
4,300	7,000 sentences	0.6 error/sentence
4,300	100,000 words	4.3 percent of sample

The last entry in the above table is again very informative, indicating that for the two technical areas under consideration the corresponding lexica are approximately 95 percent complete. It is illustrative to note that, in this connection, a computer page printout contains an average of 150-180 words, and that, from a lexical viewpoint only, the user of raw translation output is faced with a problem consisting of an error, an inaccuracy, or an untranslated term every 45-55 words. It must be noted here that a cursory post-editing process corrects, in general, only the untranslated terms.

Comparative Data Analysis

The following section, an analysis of the comparative evaluations of the raw translations produced by the lexically updated experimental SYSTRAN systems, illustrates the degree of improvement possible in lexica that are judged to be 95 percent complete.

The initial analysis of the 6,400 emendations suggested as either necessary or desirable by the subject-matter experts eliminated 2,100 as unsuitable for lexicographic updating. The remaining 4,300 were further analyzed into two large groups. Approximately 1,600 were classified as stem dictionary entries, and were entered into the stem dictionary of the experimental translation system. They were eventually used to produce the raw translations identified as Phase II translations.

The remainder, approximately 2,700 entries both of semantic expressions and of idioms, was used to update the second system dictionary and, in conjunction with the stem dictionary, was used to produce raw translations identified as Phase III translations.

The computer program developed for that purpose was used to print out sentence pairs that differed in the various raw translations. For evaluation purposes, the original translation of the 100,000 word test sample (Phase I) was compared against the Phase II output and against the Phase III output. This material was delivered to the subject matter experts for their evaluation, based on the following scheme.

A - An indication of preference for the original sentence, rather than the updated sentence. This choice could, of course, imply a change of opinion by the evaluators, or, more probably, it could mean a reduction in translation quality due to an undesirable carry-over effect of a correction in the original test sample.

B - An indication of preference for the updated sentence. This choice reflects the successful updating of a corrected sentence with the resultant expected improvement. Less likely, but still a possibility, is a secondary improvement caused again by a carry-over effect of a correction.

C - An indication of no preference between the sentences of the original and of the updated raw translations. This choice implies that the change in the two sentences was minimal enough not to be considered as affecting the sentence either in a positive or in a negative sense. This set of sentences includes those that were judged equally good or equally bad.

D - This category was introduced to reflect the viewpoint of the potential user of a raw translation. It implies that no evaluative choice could be made between the two paired sentences. Sentences were assigned to this category under the following conditions: either both sentences contained untranslated source language terms, or the Phase II sentence contained an untranslated source language term. If the original sentence contained an untranslated term, which was translated in the Phase II sentence, then the sentence was evaluated on the basis of one of the other categories.

The following tabulation depicts the evaluations of Phase I against Phase II sentence comparisons, based on the described evaluation scheme, for both technical areas and for the total test sample. The total number of sentences evaluated was 4,400, or approximately 63 percent of the total number of sentences. Values are given in percentages for both technical areas and the whole test sample.

EVALUATION TABULATION FOR TEST SAMPLE
PHASE I VERSUS PHASE II

	A	B	C	D
	Phase I Preferred	Phase II Preferred	No Preference	No Choice
Metals Working	3%	86%	9%	2%
Mechanical Engineering	1%	72%	22%	5%
Total Sample	2%	79%	15%	4%

Several facts need to be mentioned in connection with the values given in this tabulation. The first is the overall improvement in translation quality, since almost 80 percent of the affected sentences were evaluated as improved. The second is the close correlation in the overall value of category D, 4 percent, with the estimated 95 percent value, discussed in the previous section, for the completeness of the systems dictionaries. Finally, the logical expectation that the poorer of the two technical areas, mechanical engineering, would show a greater degree of improvement than metals working is not borne out. This expectation is contradicted by the "B" categories for each area - 86 percent for metals working, but only 72 percent for mechanical engineering. These values, however, do support the findings discussed for the gross error rates in the previous section.

A similar table can be constructed for the evaluation data of sentence pairs drawn from the original translation and the Phase III versions. Since both updated dictionaries are involved in the Phase III system, the number of affected sentences can be expected to increase. In actuality, approximately 4,700 sentences, or 67 percent of the total number of sentences, were evaluated during this part of the evaluation process. The following tabulation parallels that given for Phase II evaluation.

EVALUATION TABULATION FOR TEST SAMPLE
PHASE I VERSUS PHASE III

	A	B	C	D
	Phase I Preferred	Phase III Preferred	No Preference	No Choice
Metals Working	2%	92%	5%	2%
Mechanical Engineering	0.5%	79%	20%	0.3%
Total Sample	1%	86%	12%	1%

A number of observations can again be made based on the values in this tabulation. The addition of a second updated dictionary should effect the results of the evaluation in the following way. Category B, the improved sentence set, should increase. All other categories should decrease. This result is precisely what occurred. The difference between the two technical areas is still preserved, 92 percent improvement for metals working, 79 percent improvement for mechanical engineering, with a total improvement of approximately 86 percent. Perhaps the most meaningful result is the 1 percent rating for Category D in the overall sample, implying that only 1 percent of the lexical material remained untranslated after the rigorous analysis and careful updating of the two dictionaries.

The same type of evaluation process was done on the text samples translated for the carry-over effects study. The following tabulations show the results of the evaluation for this sample, which was processed in the identical fashion to the original text sample, resulting in Phase I, Phase II, and Phase III translations, and the comparison between the phases. Several differences, however, must also be noted. The first is the fact that the text samples for the carry-over effects study were modified only by the updated systems dictionaries, not initially edited by subject-matter experts. The second is the fact that the evaluation of the material differing between translations was done by a group of language experts who had not done any of the original editing. Some differences can therefore be expected in the resulting evaluation data.

EVALUATION TABULATION FOR CARRY-OVER EFFECTS SAMPLE
PHASE I VERSUS PHASE II

	A	B	C	D
	Phase I Preferred	Phase II Preferred	No Preference	No Choice
Metals Working	3%	74%	9%	14%
Mechanical Engineering	11%	64%	7%	18%
Total Sample	6%	70%	8%	16%

Categories B and D are again of special interest. Metals working registered a greater improvement, 74 percent, over mechanical engineering, 64 percent, still preserving the relative rating that these areas had received initially. Category D will be considered later. As in the study of the initial test samples, there was a Phase I - Phase III comparison and evaluation for the carry-over effects study.

EVALUATION TABULATION FOR CARRY-OVER EFFECTS SAMPLE
PHASE I VERSUS PHASE III

	A	B	C	D
	Phase I Preferred	Phase III Preferred	No Preference	No Choice
Metals Working	8%	54%	23%	15%
Mechanical Engineering	4%	71%	7%	18%
Total Sample	7%	61%	16%	16%

The results of this evaluation do not continue the trend exhibited in all the other tables thus far analyzed. The expected trend, decrease in Categories A, C, and D, and a corresponding increase in Category B over the Phase II tabulation values, holds only for the results for mechanical engineering. One immediate explanation for this result is statistical. One of the evaluations for metals working exhibited completely different results from the others. The calculated percentages without that one evaluation read as follows:

	A	B	C	D
Metals Working	10%	70%	3%	17%

The modified percentages for the total sample, using all the data except this one single set results in the following:

	A	B	C	D
Total Sample	6%	71%	5%	18%

Although these new percentages are closer to the expected trend, they are not entirely correlated with the expected results, which were achieved with the remainder of the evaluation data. There may, therefore, be additional reasons for this discrepancy. These will be given in the following section. The following tabulations illustrate both the general trend discussed above for the several technical areas and the problem with the one data set. The parentheses around one set of values indicate the statistically modified values.

METALS WORKING

Category	Initial Sample Phases		Carry-Over Sample Phases		
	I-II	I-III	I-II	I-III	
A	3%	2%	3%	8%	(10%)
B	86%	92%	74%	54%	(70%)
C	9%	5%	9%	23%	(3%)
D	2%	2%	14%	15%	(17%)

MECHANICAL ENGINEERING

Category	Initial Sample Phases		Carry-Over Sample Phases	
	I-II	I-III	I-II	I-III
A	1%	0.5%	11%	4%
B	72%	79%	64%	71%
C	22%	20%	7%	7%
D	5%	0.3%	18%	18%

TOTAL DATA

Category	Initial Sample Phases		Carry-Over Sample Phases		
	I-III	I-III	I-II	I-III	
A	2%	1%	6%	7%	(6%)
B	79%	86%	70%	61%	(71%)
C	15%	12%	8%	16%	(5%)
D	4%	1%	16%	16%	(18%)

The tabulation for total data percentages leads to a further brief discussion of the significance of Category D. The values derived for Category D are strongly indicative of an additional point that needs to be stressed. The editing of the initial data sample and the subsequent updating of dictionaries resulted in a reduction of untranslated or unfound terms. In the carry-over effects study, however, the initial Category D values were higher than those for the original study, and they showed no reduction in Phase II and Phase III translations. This result strongly implies both that further study of the lexica is needed and that further improvement in translations can be achieved through additional sample analysis and dictionary updating. The following tabulation lists the Category D values for the discussed technical areas as an illustration of this point.

CATEGORY D TABULATION

	Initial Sample Phases		Carry-Over Sample Phases	
	I-II	I-III	I-II	I-III
Metals Working	2%	2%	14%	15%
Mechanical Eng.	5%	0.3%	18%	18%
Total Sample	4%	1%	16%	16%

The data thus far presented and discussed reflect the evaluation into four categories of the set of sentences that differed in two related translations. The conclusions drawn from the data do not, therefore, reflect the effects of the updating of system dictionaries on the complete data sample. The following discussion covers this specific area.

Of the approximately 7,000 sentences contained in the initial test sample and the updated Phase II translation, 63 percent were affected by the changes in the stem dictionary, either directly through the corrections made during the editing process, or through carry-over effects within the initial test sample. Of the number of changed sentences, 79 percent fell into Category B; i.e., they were judged to be better than the original. These two percentage figures combine to result in an overall improvement of approximately 50 percent. That is, fully half of the sentences in the test sample were improved by the lexical updating of the stem dictionary.

Since a large number of the lexical corrections were incorporated into the semantic expressions and idioms dictionary, it is essential to consider also the degree of improvement in translation quality when the source material was translated against both updated dictionaries. In this second case the number of affected sentences, when the original text was compared with the Phase III translation, increased to 67 percent. Of this number of changed sentences, 86 percent were judged to be better than the original sentences, resulting in an overall improvement of 56 percent of the number of sentences in the test sample. It must be stressed that in both of these cases the evaluation of the updated sentences was done by the same bilingual subject-matter experts who had edited the original source language text. The following tabulation illustrates these facts.

INITIAL TRANSLATION VERSUS PHASE II AND PHASE III

	Initial Translation versus Phase II	Initial Translation versus Phase III
Sentences Affected	63%	67%
Sentences Improved	79%	86%
Overall Improvement	50%	56%

The values given in the above tabulation again illustrate several important points. One is the real and obvious improvement achievable in translation quality through an efficient updating of the stem dictionary of a machine translation system. The second point is the 6 percent increase in improvement in translation quality when the semantic expression dictionary was updated as well. Since, however, of the approximately 4,300 corrections made by the subject-matter experts and used for dictionary updating, 2,700 corrections were incorporated into the semantic expressions dictionary, and only 1,600 corrections were used to update the stem dictionary, the 6 percent improvement due to 2,700 entries seems disproportionately small when compared with the 50 percent improvement achieved by 1,600 entries. These figures are somewhat misleading, since there exists a definite overlap between Phase II and Phase III translations. That is, the updated semantic expressions dictionary affected both sentences that had previously been affected by the updated stem dictionary and sentences that had not previously been affected. The overlap between Phase II and Phase III sentences is considerable. Tests were run on Phase II and Phase III translations of the original test material. Although the results of these comparisons were not evaluated, they did show that 42 percent of the updated Phase II sentences were further affected by Phase III changes. The 6 percent quantitative improvement does not actually reflect the qualitative improvement due to the updated semantic expression and idioms dictionary.

The same process was applied to the source language text sample for the carry-over effects study. It was expected that the dictionary updating would extend beyond the localized context of the originally edited text sample. The carry-over effects study was to discover the global effects of the updating when applied to related source language material. In general terms, since it is practically impossible to match the degree of difficulty of natural language texts exactly, the expected

result was a decrease in the number of sentences judged better in the Phase II and Phase III translations when compared with the result of the initial analysis. In addition, the raw translations of this related material obtained from the two experimental SYSTRAN systems were compared and evaluated by a set of evaluators different from the initial editors. A further decrease was therefore expected due to the use of different evaluators, although this use, of course, also had a definite advantage. It tended to eliminate any possible bias in evaluation introduced, consciously or unconsciously, by editors evaluating their own work. Nevertheless, the findings of the carry-over effects study showed some very positive results. The following tabulation parallels the one given for the original study and is based on the 6,400 sentences of the carry-over effects test sample.

CARRY-OVER EFFECTS STUDY

	Initial Translation versus Phase II	Initial Translation versus Phase III
Sentences Affected	57%	67%
Sentences Improved	70%	61%
Overall Improvement	40%	41%

In general, the results of this study confirmed the expected outcome. As in the initial study, there was an increase in the number of sentences affected when going from Phase II to Phase III translations. Paralleling the increase in the number of sentences, there is an increase in the number of improved sentences in the overall category. The decrease in improvement in going from Phase II sentences to Phase III sentences can, in part, be attributed to the output of one evaluator in this study, as has been discussed previously. Percentages for the original versus Phase III sentences, when the results of this one evaluator were excluded, were 71 percent for improved sentences, and 47 percent for overall improvement. Although the amount of improvement decreased in the carry-over effects study, there is still a considerable improvement in translation quality.

The values presented in the tabulations in this section, the relationships developed between these values, and the results thus far derived will be further discussed and analyzed in the following section.

SECTION IV

CONCLUSIONS AND RECOMMENDATIONS

Introduction

The interpretation, analysis, and correlation of the experimental data, obtained through the evaluation of the raw translations of Phase I, and the related data gathered from the evaluation of Phase II and Phase III translations, both of the initial material and of the carry-over effects study, led to the conclusions covered in the previous section. These conclusions were limited, to a certain extent, to the context, first, of the specific translations of the source language test material, and, second, also limited to the constrained area of strictly lexical updating of the two technical areas covered by the test material.

There are, however, several additional areas that need to be discussed. In part these areas are also the result of the lexicographic work done during the current research, but they extend beyond the scope of the restricted lexicographic updating. These areas need to be analyzed, therefore, for their possible impact on the SYSTRAN system as an integrated whole.

The following discussion is divided into four general sections: the first of these is a further generalized analysis of the results and the relationships between the initial study and the carry-over effects study of the source language test material; the second section is a discussion of evaluation methodology as developed from the evaluation results; the third section covers conclusions and recommendations, based on the findings of this research, but restricted to the specific area of lexicographic analysis and updating and the specific methodology used to implement this updating; the final section covers recommendations concerning the entire SYSTRAN machine translation system, of which the updated dictionaries are only one important component. These recommendations are all based on the data gathered during the analysis of the test material, and the results obtained from this analysis.

Discussion of Problem Areas

The following tabulations repeat, for easy reference, the values given in several previously listed tabulations. They also contain some additional information, contrasting the results of comparisons among the different translation phases.

INITIAL STUDY
(Sample Size: 7,000 sentences)

	Original Text versus Phase II	Original Text versus Phase III	Percentage Increase
Sentences Affected	63%	67%	6%
Sentences Improved	79%	86%	8%
Total Improvement	50%	56%	12%

CARRY-OVER EFFECTS STUDY
(Sample Size: 6,400 sentences)

	Test Text versus Phase II	Test Text versus Phase III	Percentage Increase
Sentences Affected	57%	67%	17%
Sentences Improved	70%	61%	-13%
Total Improvement	40%	41%	2%

The results of the lexicographic updating of the system stem dictionary, based on the use of 1,600 entries from a total of 4,300 suggested emendations, demonstrated clearly the improvement achievable through this process. The achieved improvement due to the lexicographic updating of the system semantic expression dictionary, based on the remaining 2,700 emendations, seemed disproportionately small when contrasted with the improvement due to the updated stem dictionary. A second problem was the actual decrease in the number of improved sentences in the carry-over effects study when passing from Phase II to Phase III translations, although there was an overall improvement due to increase in the total number of sentences affected.

Two important parameters of a natural language translation need to be considered in this context, translation completeness and

translation accuracy. These two parameters together influence the quality of a translation, although they are difficult to gauge individually. In general terms, however, it can be stated that stem dictionaries consisting of single lexical units contribute more to translation completeness and general accuracy in terms of the amount of material translated, whereas semantic expression dictionaries contribute more to specific, contextual accuracy rather than to translation completeness.

In general, then, it can be expected that the introduction of a semantic expression dictionary will lead to the improvement of some sentences that had already been affected by an updated stem dictionary, as well as to the improvement of some sentences not previously affected. It is this overlap that leads to the small overall increase in improvement. In actual fact, the overlap between the Phase II and Phase III sets of sentences was 42 percent for the initial study, or approximately 2,900 sentences. These values indicate the considerable impact of a semantic expressions dictionary on overall translation quality. This overlap can, therefore, be considered as one of the factors limiting the total number of improved sentences. Another factor, also limiting the increase in improvement to 8 percent only is the following: whenever possible, technical personnel of the Technical Translation Group in Dayton introduced all corrections submitted for their consideration into the lexica of the system. These corrections included some that were felt to be of very limited applicability because of extremely restricted contextual meanings; they also included some corrections that would usually have been incorporated into other components of the SYSTRAN system. All of these corrections were introduced into the lexicon because of the restriction of this research to work with the system dictionaries and because they could marginally be analyzed as lexical entries.

The above two reasons can be considered valid reasons for the limited 8 percent increase in the number of sentences improved for the initial study; they can account only in part for the decrease in the number of improved sentences for the carry-over effects study. That is, the overlap between these sets of sentences may be much smaller than in the initial study. No complete test was made, but sampling checks indicated an approximate overlap of less than 25 percent. Similarly, the limited value of some of the semantic changes coupled with the general restrictions of such expressions to translation accuracy rather than translation completeness could also have contributed to this result. The source language test material could have differed enough from the initially analyzed material to cause no carry-over or erroneous carry-over. (The impact of the results of the evaluation of one translation analyst has already been discussed.) However, no definite reason for the decrease could be established.

A further question that can arise in considering the improvement achieved in these tests is a statistical one: can the improvement achieved at any stage of the tests be attributed to chance effects? Obviously, the 50 percent and 40 percent improvements achieved in the initial and in the carry-over effects study, respectively, through the stem dictionary updating are considerable enough so that they need not be considered in this context. Although statistical studies on natural language samples have to be interpreted with care, due to the very nature of the data, it is informative to consider the results other than the outcome of the stem dictionary updating from a statistical viewpoint.

The individual sentences in the test samples can be considered as discrete, independent objects, since they are translated with no reference to each other. Since each of them also has the possibility either of being improved or of not being improved in a given test stage, a modified approach to the probabilities of the binomial distribution can be used to determine the possible influence of chance effects. A simple example will illustrate this procedure.

Let $n = 10$ be a given sample size, and let $x = 5$ be the number of successes at some trial. Tables exist for the two-sided limits of the binomial distribution. [See, for example, A. Hand, Statistical Tables and Formulas.] For $x = 5$, and $n - x = 5$, at a 95 percent confidence level, the limits are 0.187 and 0.813. These values imply that on successive tries any value for x, $2 < x < 8$, could be expected for the given sample of $n = 10$. However, values of x other than the 2 - 8 range cannot be entirely attributed to chance effects. This method will be applied to the language studies performed on this project, with the following modification. Computed statistical tables are limited in size and do not go high enough for sample sizes of 7,000. For such large values of n, however, the binomial distribution converges to the normal distribution, hence approximation formulas exist to calculate values corresponding to the explicit tabular confidence limits.

Let n be the given sample size, and let p be the probability for improving a given sentence, and let $q = 1 - p$. Then the required limits are given by $p \pm 2s$, again for a 95 percent confidence level, where $s = \sqrt{pq/n}$. For the cited example of $n = 10$, the calculated values of 0.184 and 0.816 compare favorably with the tabular values of 0.187 and 0.813. The following tabulation presents the statistical findings for the different sets of test translations.

BINOMIAL DISTRIBUTION LIMITS
95 PERCENT CONFIDENCE LEVEL

Translations	Improvement Probability	s Value	Two-Sided Limits (approximate)
Original, Phase II	0.5	0.00598	0.494 - 0.506
Original, Phase III	0.56	0.0051	0.555 - 0.565
Carry-Over, Phase II	0.4	0.00612	0.394 - 0.406
Carry-Over, Phase III	0.41	0.00618	0.404 - 0.416

Statistical results obtained from natural language samples must always be interpreted with a certain amount of care. Nevertheless, in this case it is both interesting and informative to look for statistical confirmation of already obtained results. Obviously, the results presented in the above tabulation can be used to test for chance factors in translation improvement. For strict interpretation, an entry in the tabulation should be compared with successive results from a similar test; i.e., a Phase II translation of the original test material should be compared with the results of another translation of the same type. However, there exists another possible interpretation of the calculated statistical limits when two different samples are being compared; chance factors can be regarded as minimal if the calculated limits do not overlap. On this basis the above tabulation shows only one questionable area, the degree of improvement achieved in going from a Phase II to a Phase III translation of the carry-over effects study. The upper limit of the Phase II sample and the lower limit of the Phase III sample overlap, 0.406 and 0.404, indicating the possible presence of chance factors, and further confirming the questionable results of this previously discussed problem. Of equal importance is the fact that there is no overlap in the calculated limits when corresponding Phase II or Phase III samples between the initial and the carry-over effects studies are compared. These results, in turn, support the conclusion that the improvements achieved in translation quality and in carry-over effects due to the lexicographic updating are due mainly to the analysis and the updating performed during the course of this research effort.

Discussion of Evaluation Methodology

A general question of importance that must be considered in connection with the obtained results is that of evaluation of machine-aided translations. Two preliminary points must be made. As had been discussed previously, no existing, off-the-shelf evaluation method is entirely satisfactory. The main drawback of all of these methods is the fact that, although some of these methods do demonstrate positive correlation between translation acceptability and scores obtained from certain measurable factors considered in these methods, none of these methods really measures the one factor that is of paramount importance - translation correctness. The one method that does evaluate translations on that basis is the method that uses an expert bilingual for the evaluation. The disadvantage of this approach is, of course, the degree of subjectivity that may be introduced into such an evaluation by the bilingual expert.

The second point to be considered is not merely how the evaluation of a translation is performed but also what it is that is being evaluated. A raw translation produced by a machine translation system can be evaluated as a representative, unrestricted sample of natural language and evaluated as such, or it can be considered as output from a restricted, clearly defined system and evaluated as a sample of natural language constrained within specific and known system design parameters. For machine translations of technical material a further consideration can be listed, and that is the eventual usefulness of the raw translation to the potential user, who usually has no knowledge of the source language itself. Given these points, it is to be expected that an evaluation by a bilingual expert will be rated higher for a machine-aided translation interpreted as output from a machine translation system than if it is considered as an example of the unrestricted natural language. This assumption can again be confirmed from results of the current research. The two technical areas considered in this research were mechanical engineering and metals working. These areas were rated by Technical Translation Group technical personnel as "below average" and "average", respectively, based on a five-point scale ranging from "good" to "poor". These evaluations were made by bilinguals, knowledgeable in machine translation work, and expert in the specific machine translation system, SYSTRAN. Their evaluations are, therefore, a relative rating, based on at least three factors: knowledge of the system limitations, knowledge of the source language, and comparison of the translations in mechanical engineering and metals working with other technical areas presently processed by SYSTRAN.

Evaluation of the raw translations in the same two technical areas by the bilingual subject-matter experts involved in this project

rated the two areas as "D" for metals working and "between D and F" for mechanical engineering. These evaluations were based on the usual A through F five-point scholastic scale, together with the following factors: knowledge of the source language, expertise in the appropriate technical subject matter, consideration of usefulness of the product to the potential user, and general background knowledge of machine translation, but no specific knowledge of the SYSTRAN system.

Since any computer-based system has certain built-in limitations due to individual computer design, and since a natural language translation system contains a certain number of linguistic limitations as a consequence of basic system design parameters, it follows that a machine translation evaluation method should be based on the theoretical capabilities of the system. This type of evaluation has as a consequence the requirement that the evaluation of raw output should be done by bilingual experts, preferably with the technical knowledge necessary to evaluate the technical accuracy of the translation, but certainly with the system knowledge necessary to evaluate such raw output in terms of the limitations and capabilities of the system.

The question of subjectivity in the evaluation process must also be considered. The analysis of some of the data obtained during this research demonstrated that careful evaluation by bilingual experts correlated only to a certain extent with results obtained through the method used throughout this research. Until an objective, reliable, and valid method for natural language evaluation is developed, the use of bilingual experts must remain the preferred method, in spite of its inherent dangers.

Although the degree of improvement was established during this research, both for the initial and for the carry-over effects study in two technical areas, no thorough evaluation was done on the complete samples in the sense of assigning a rating to the improved samples. An approximate evaluation was done by the bilingual subject-matter experts and the Phase III translations from both of the technical areas rated at least one point higher on the five-point A - F scale than the original raw translation. No corresponding confirmatory evaluation was performed by technical personnel of the Technical Translation Group. However, the current research suggested a possible aid to bilingual evaluators in the area of consistency and limited objectivity.

Consistency in evaluation is aided by a clearly defined scale to help bilingual evaluators eliminate the subjective in their work. The following approach is offered as a theoretical possibility for such

a scale. Again, it is based on the results of the research into the effects of lexicographic updating. The area of mechanical engineering is rated poorer in translation quality than that of metals working. Correspondingly, the amount of improvement achieved in metals working was higher, 92 percent, than in mechanical engineering, 79 percent. With additional and similar research in both of these areas and all other areas processed by SYSTRAN, it should be possible to develop error or percentage ranges for assigning specific ratings to a given translation based on the amount of improvement achievable through a careful editing process. Such an approach would develop an objective scale for assigning evaluation rates to raw translations. The current limited research results demonstrate a trend toward a possible useful correlation between translation quality and possible improvement.

Conclusions and Recommendations for the Lexical Component of the SYSTRAN Machine Translation System

The main purpose of an operational machine translation system is the production of useful and technically correct translations for actual and potential users of the system, and the continuous improvement of the quality of these translations whenever possible. Since it has been demonstrated that a concentrated effort in lexical analysis and subsequent updating of systems dictionaries can lead to a significant improvement in system output, the third area of discussion addresses itself to suggestions, recommendations, and conclusions in lexicographic work.

This area of discussion will treat three separate, though related, aspects of lexicographic analysis as performed during the current research. The first is a logical extension of the present approach, limited in application to the two technical areas thus far discussed; the second is a series of conclusions that can be drawn from the current approach and its extension; and the third is a further extension of the method to the entire area of lexicographic work, as defined within the SYSTRAN system.

The method used to evaluate and to improve the lexicon of the two test areas is an iterative method. It is iterative in the sense that the same test material is submitted to repeated translation by an updated translation system, with a corresponding evaluation of the new translation relative to the old one. This process can, of course, simply be extended further. The present iterative process passes through three phases, from the Phase I production and evaluation of the original

test material to the production and evaluation of Phase III material. This whole process, however, could be considered as one single iteration. At this point, after evaluation, the Phase III translation could be redefined as the initial step in the evaluation process. That is, this translation could now be submitted to the bilingual subject-matter experts for the same type of careful lexical analysis that they performed on the initial Phase I test material. The emendations made to this Phase III material could then again be analyzed for insertion into stem and semantic expression dictionary entries, the respective dictionaries could be updated, and new Phase II and Phase III translations could be generated. The existing computer program could again be used to compare corresponding translations for evaluation of the achieved improvement during this second cycle.

This step of replacing Phase I material with the improved Phase III material, and repeating the evaluation and updating process for the complete cycle, could then be repeated any desired number of times until no further improvement could be discerned in the technical lexica under consideration. This extension of the present method would require some modifications in the computer program developed for this project. The program would have to print out additional information on sentences being compared to assist evaluators with sentences improved during preceding cycles. No other changes or modifications would be needed in the present method to implement these additional steps. Careful attention would have to be paid to termination criteria for such a process to avoid a possible oscillation due to repeated duplicated lexical changes. In addition, the true contribution of both stem and semantic expression dictionaries to translation quality could be established more accurately, especially in the area of overlap.

Together with this extension of the method for updating lexica, it is possible to draw a certain number of conclusions, still based on the results of this project, in terms of the two factors of raw translation improvement and cost-effectiveness of the method.

Translation quality can be improved to a considerable degree through careful and thorough lexical analysis of selected, representative test material. A major advantage of this approach to the improvement of translation quality of an operational machine translation system is the fact that no major changes are necessary in the basic design of the system. The dictionaries are already an integral component of the system and the updating of the dictionaries can be accomplished through existing methods and programs.

The updating of system dictionaries produces strong carry-over effects from the test samples to related technical texts. The rate of improvement appears to be a diminishing one, although the actual rate is not identifiable from the results of the present research. This diminishing rate could be determined through additional carry-over effect studies using both the extended method described above and additional sets of test material in the same and in related technical areas. Such a study could be conducted either independently or as part of the extended iterative approach to lexical evaluation.

Of extreme practical importance, however, is the conclusion reached on the cost-effectiveness of this process. In general, the cost-effectiveness of the lexicographic updating is due mainly to the carry-over effects of the lexical changes, with the resulting amortization possibilities of the cost of the initial changes over an extended period of time. An additional contributing factor in this cost-effectiveness area is the effect that a lexical improvement of 50 percent, as achieved in this research, must have on raw output acceptability, and the consequent effect of this acceptability on post-editing processes. Such an improvement should reduce, to a great extent, both the need and the number of requests for post-editing processes. (A more detailed analysis of the cost-effectiveness of lexicographic updating is presented in the report, "An Experimental Measure of the Cost-Effectiveness of Expanding SYSTRAN Subject-Area Dictionaries", dated August 17, 1976, presented in Appendix C.

Based on the same factors of translation quality and cost-effectiveness, coupled with past experiences of the Technical Translation Group of the Foreign Technology Division with SYSTRAN, a number of additional recommendations, related to the work thus far done, can also be made. These recommendations reflect further logical applications of the present method.

The SYSTRAN machine translation system contains twelve technical lexica in the system dictionaries. These dictionaries are rated by Technical Translation Group personnel according to the quality of raw translation produced by them. Two technical areas, mathematics and physics, are rated "good"; one, aviation/space, is "above average", and one, mechanical engineering, is "below average". The remaining eight are rated "average": a general dictionary, electronics, computer technology, biology/medicine, military science, chemistry, earth science, and metallurgy, which includes the area of metals working. The same type of study that was conducted for the two areas of mechanical engineering and metals working can be conducted for the other technical areas with the same degree of linguistic difficulty; e.g., biology/medicine. The same degree of improvement that was achieved for the two technical areas used in this research can probably be achieved for these areas as well.

Finally, further careful research in this type of lexicographic updating, concentrated in one of the technical areas like physics, rated "good" in terms of machine translation output, would help to determine, first, the immediate amount of improvement achievable in such an area, as contrasted with one that is poorer in terms of translation quality. Second, such a comparison could also lead to two additional developments, the possibility of establishing quickly and easily ranges for the theoretical evaluation scale discussed previously, and also the determination of an upper bound for possible improvement through lexicographic updating.

Conclusions and Recommendations for the Integrated SYSTRAN Machine Translation Systems

The recommendations of the previous section on the lexicographic research led to the final area of consideration. SYSTRAN models, to a certain extent, a natural language system and contains more components than just a lexicon. There exist also semantic and syntactic components, contributing to the accuracy of a raw translation. It follows, therefore, that although a significant amount of translation improvement can be achieved through the use of the lexicon, there is a limit to lexical improvement. The following brief discussion of SYSTRAN as a linguistic system is still based on results of the current research. Specifically, it is based on the analysis of the total number of emendations made by the subject-matter experts. Of the approximately 6,400 suggested corrections, 4,300 were actually used for the lexicographic updating process. The remainder, 2,100 corrections, was rejected by Technical Translation Group analysts as unsuitable for purely lexicographic work as it applies to SYSTRAN. However, this set of rejected corrections can in its turn be further analyzed and separated into several large groups as follows.

Typographical Errors. The large majority of these errors was attributable to key-punching errors. Some were due to misinterpretation of the original source language text, and a few were due to actual errors in the source language text itself. The percentage of these errors, based on the total number of rejected corrections, was approximately 14 percent.

Corrections to Lexical Routines. Lexical routines are an extension of the SYSTRAN lexicon. Modifications to these routines fell outside the scope of the current research. Corrections that required

either modification of existing routines or development of new routines therefore were not incorporated into the system. The percentage of corrections fitting into this category was approximately 10 percent.

Modifications of Existing Software. The corrections to the SYSTRAN system under this category involved moderate changes to existing modules of the system. These changes again fell outside the scope of lexical analysis, and included such things as prepositional conflicts, proper names, and measures and weights terminology. The percentage for this category was approximately 27 percent.

Extensions to Existing Software. This category included those corrections that could still be incorporated into the SYSTRAN system; these would require an extended effort to implement and to test, since they involved either the development of extensions to existing programs or the development of new ones. Generally, these corrections fell into the areas of rearrangement routines, developing correct English syntax and word order from Russian word strings. The percentage for this category was approximately 12 percent.

Corrections Impossible To Implement. This category was approximately 7 percent of the total. It included corrections that would have involved basic changes to the design of the SYSTRAN system. An example of this type of correction is one that required analysis of the context of more than one sentence. SYSTRAN, however, can only translate on a sentence-by-sentence basis.

Corrections Analyzed, But Not Implemented. This category, comprising approximately 30 percent of the total, was the miscellaneous class. Included in this category were corrections that fell into such varied areas as unexplained system failures; suggestions made by the subject-matter experts that required additional, but unobtained, clarification; and corrections that were only stylistic variants of the original correct translation.

This type of negative data gathered from the lexical analysis produced a number of corrections in the category described as "Extensions to Existing Software" that were considered valuable enough to place into a separate subcategory for future analysis and possible incorporation into the system. The same data can also be used for a general overview of the SYSTRAN system as an integrated whole, with possible improvement to its components, and to its product, an acceptable technical translation through the interaction of these components.

Two approaches can be taken in the development of such an integrated systems analysis. The first is the study of the entire system with all of its components and their interaction considered simultaneously. The second consists of the analysis of the effect of a perturbation in one component on the remainder of the system. Frequently, however, an additional factor must be taken into consideration. That is the presence of two complex systems acting together, as is the case with SYSTRAN, which consists of a primary linguistic system superimposed on a supportive computer system, each influencing the efficiency of the other. In specific cases, for a thorough system integration analysis, the existing man-machine interface must also be considered in the work toward optimization of the entire system.

The main effort of the current machine translation research was directed at an analysis and subsequent evaluation of one portion of the SYSTRAN system, its lexical component, with the major aim of improving translation quality through improvement in the lexicon. The following recommendations developed from a system integration analysis are, therefore, aimed, first, at the purpose of further continued improvement of the lexical component, and, second, on an analysis of the possible effects of this improvement on the remaining major components of the linguistic system, together with independent areas of analysis in these other components. Some attention will be paid to the man-machine interface areas of the SYSTRAN system. These areas include the input operations and the various operations connected with the raw translation output. The effect of these changes on the efficiency of the supportive computer system, however, is expected to be of such a minor nature that it can be omitted in these considerations, except in one critical case.

These recommendations for the linguistic portion of the SYSTRAN system are arranged in the order of ease and speed of possible implementation, with considerations both of immediate and of long-range benefits. The recommendations cover the three major components – lexical, syntactic, and semantic – of the linguistic system, and the appropriate parts of peripheral programs. It must again be stressed that all of these recommendations are made within the constraints of the design specifications of SYSTRAN.

Linguistic Component of SYSTRAN

Lexicon

The suggested improvements in the lexical component have already been covered in preceding sections, in terms of the twelve technical areas presently translated by SYSTRAN, and in terms of the stem and semantic expression dictionaries into which these technical glossaries are arranged. An immediate result of the current research is a strong recommendation that an extensive effort be made for a similar concentrated analysis of the remainder of the technical glossaries, together with a thorough study of the relative contributions of the stem and semantic expression dictionaries. Since a subjective rating of these glossaries already exists, this evaluative effort should proceed from the poorer to the better cases, giving the possibility of stopping the evaluative process if returns indicate decreasing resultant benefits. Based on the findings of the current research, it is reasonable to assume that this additional research would lead to improved translation quality in terms of translation completeness at practically no reduction in the efficiency of the computer component of the SYSTRAN system. Obviously, this work with the lexical component would not necessitate changes in the design of SYSTRAN.

Lexical Routines

The SYSTRAN machine translation system contains a set of lexical routines, accessed by the system through its own macrolanguage, on the basis of certain key words. These routines are used mainly to resolve ambiguities in running text. Although the analysis and use of lexical routines exceeded the scope of the current research, a certain amount of the data gathered through the analysis of the suggested changes during the lexical evaluation process fell directly into the area amenable to correction or further improvement through the use of lexical routines. It follows, therefore, that the implementation of changes or additions to lexical routines supported by the data gathered for mechanical engineering and metals working would result immediately in further improvement in translation quality for these areas. Improvement for the entire system could be achieved through additional data applicable to lexical routines; the data could be obtained from the implementation of the recommendation for improvement of the entire lexicon. In the area of lexical routines, an improvement in translation quality can be obtained with little additional cost to the supportive computer system, since both the concept and the general specifications for the lexical routines, as well as the macrolanguage for their use in the system, already exist.

Semantics

One area of possible research that could lead to an improvement in translation quality in terms of translation accuracy is that of semantics, as defined for the SYSTRAN system. The lexicon in SYSTRAN incorporates a number of semantic codes attached to individual lexical entries. The recommendation in this area is for a separate, concentrated study of both the existing set of semantic codes for general applicability and for logical completeness, and an analysis of these codes in terms of possible optimal internal organization, e.g., hierarchical, inclusive, etc. This recommendation is obviously not for immediate but for long-range results. It is not directly related to lexical analysis, in terms of the current research, although its implementation would affect both the dictionaries and other system components. The eventual benefits from a thorough analysis of this semantic area would again result in translation improvement in terms of accuracy. Although the implementation of possible further recommendations arising from such a study would be reasonably difficult, these recommendations would still be based on already existing concepts and components of SYSTRAN, with little effect on computer system efficiency.

Syntax

Because of the difficulty in implementing the recommendation covered under this heading, it is listed as the last point in the sequence of suggestions for linguistic system improvement, although the benefits derivable from its implementation are expected to be considerable. In general, the problem consists of analyzing a given sequence of text of the source language, not in terms of its lexical components but in terms of universal syntactic components, as given by a phrase structure representation. The second part of the problem is then to rearrange this structure to conform to the corresponding structure in the target language. Again, the basis for this approach already exists in the SYSTRAN system. The routine designed for this purpose does need to be integrated into the system using both the information given by the lexicon and the grammatical analysis of strings given by other components of the existing system, such as the existing and operational set of syntactic routines. The general specifications for such a rearrangement routine also exist for the SYSTRAN system, although efficient implementation would require an extended effort in linguistic analysis and programming and systems study. This recommendation, if implemented, would also have an appreciable effect on computer operations, since it can be reasonably expected that phrase structure analysis and rearrangement would be

time-consuming in terms of computer usage, even assuming efficient tree-mapping algorithms. The effect on translation quality, however, would be appreciable, with a considerable carry-over effect into the area of one aspect of man-machine interface.

Man-Machine Interface

SYSTRAN Input

Several factors need to be mentioned briefly in connection with the preparation of input for SYSTRAN. These include the amount of input that can be processed by a staff limited in size, and the error rate attributable to human operators with no knowledge of the source language. Obviously, these problems can be improved by appropriate staff increases and increased attention to quality control in input preparation. The long-range solution, both for increase in input and decrease of error rate, is the development and eventual use of appropriate OCR equipment for direct input operations.

Output Analysis

SYSTRAN produces a raw translation for the use of the system user. Frequently, the quality of the raw translation requires a post-editing process to produce an improved version known as a preliminary edited translation. A second degree of post-editing, resulting in a finished machine translation, has been eliminated in favor of direct human translation without recourse to the computer translation system. The implementation of the recommendations for the lexicon, the lexical routines, and the semantic components of SYSTRAN would result in improved raw translations, with less need for preliminary edited translations. Similarly, the addition of a syntactic rearrangement routine could be expected to reduce the need for finished machine translations or corresponding translations by human experts.

An interpretive model of a natural language translation system can be represented in general terms as consisting of three major components, lexical, syntactic, and semantic. This representation covers only the linguistic portion of the system, of which the other portion is the supportive computer system. Both of these parts of the system can be subjected to systems analysis for optimization purposes. The computer system, on which the linguistic system is superimposed, can very easily be analyzed for greatest possible efficiency in terms of computer

utilization, but it should not be done at a loss of completeness and accuracy of the linguistic system, since the main purpose of the combined systems is the production of acceptable natural language output.

Because of this primary purpose of the SYSTRAN machine translation system, efficient computer usage has not been covered in this system integration analysis. The peripheral input and output operations have been covered briefly. For the main components of the linguistic system, recommendations have been offered for lexical, syntactic, and semantic improvements to be developed and integrated into the existing system. These recommendations are submitted for possible future consideration. They are based on the findings of the research into the lexical component, restricted by the basic design of SYSTRAN itself, and offered with a reasonable expectation for successful implementation and for overall improvement of translation quality.

APPENDIX A

BIBLIOGRAPHY

1. Arthur D. Little, Inc., "An Evaluation of Machine-Aided Translation Activities at FTD", 1965.

2. Carroll, J. B., "An Experiment in Evaluating the Quality of Translations", Mechanical Translation and Computational Linguistics, 1966, Vol 9 (3 & 4), pp 55-66.

3. Crook, Mason N., and Bishop, Harold P., "Evaluation of Machine Translation", Final Report, The Institute for Psychological Research, Tufts University, April, 1965.

4. Darnell, Donald K., "The Development of an English Language Proficiency Test of Foreign Students, Using a Clozentropy Procedure", Final Report, Project No. 7-8-010, Grant OEG-8-8-070010-200(057), October, 1968.

5. Leavitt, Alvan W., Gates, Jesse L., and Shannon, Susan C.,"Machine Translation Quality and Production Process Evaluation", RADC-TR-71-206, October, 1971.AD#732886.

6. Miller, G. A., and Beebe-Center, J. G., "Some Psychological Methods for Evaluating the Quality of Translations", Mechanical Translation, 1958, Vol 3, pp 73-80.

7. Nakamura, Yukio, "Identification of Languages With Short Sample Texts, A Linguometric Study", Library and Information Sciences, No. 9 (1971), pp 459-481.

8. Orr, D. B., and Small, V. H., "Comprehensibility of Machine-Aided Translations of Russian Scientific Documents", Mechanical Translation and Computational Linguistics, 1967, Vol 10, pp 1-10.

9. Pfafflin, Sheila M., "Evaluation of Machine Translations by Reading Comprehension Tests and Subjective Judgments", Mechanical Translation, 1965, Vol 8, pp 2-8.

10. Shannon, Claude E., and Weaver, Warren, "The Mathematical Theory of Communication", The University of Illinois Press, Urbana, Ill., 1949.

11. Shannon, C. E., "Prediction and Entropy of Printed English", Bell System Technical Journal, Vol. 30, 1951, pp 50-64.

12. Sinaiko, H. Wallace, and Klare, George R., "Further Experiments in Language Translation: Readability of Computer Translations", Institute for Defense Analyses, Arlington, Va., August, 1971.

13. Sinaiko, H. Wallace, and Klare, George R., "Further Experiments in Language Translation: Readability of Computer Translations", Institute for Defense Analyses, Arlington, Va., December, 1971.

14. Spolsky, Bernard, Bengt Sigur, Masahito Sato, Edward Walker, and Catherine Arterburn, "Preliminary Studies in the Development of Techniques for Testing Overall Second Language Proficiency", Language Learning, Special Issue No. 3, 1968, pp 79-101.

15. Szanser, A. J., "Machine Translation - The Evaluation of an Experiment", The Incorporated Linguist, October, 1967, Vol 6, pp 90-95.

16. Taylor, Wilson S., "Cloze Procedure: A New Tool for Measuring Readability", The Journalism Quarterly, Fall, 1953, pp 415-433.

APPENDIX B

SYSDIF COMPUTER PROGRAM DESCRIPTION

The SYSDIF Program compares sentences from two SYSTRAN system print files. The print files are generated by translating the same material using two different versions of the SYSTRAN system. These print files of translated material are matched sentence by sentence to see if there has been a change in sentences of one translation as compared with the other translation. Sentences that are different from one translation to the other are listed by the program. The program processing flow is given in Figure 1.

The program is written in FORTRAN EXTENDED. It is operational on Battelle's CDC CYBER 73 computer. The listing of the FORTRAN EXTENDED source code along with SCOPE operating control cards accompanies this report. The program required 850 system seconds on the CDC CYBER 73 to run the 10,000-word comparison of two translations.

Inputs to the SYSDIF program consist of two tape files. The tapes are the output print tapes of the same material translated by two different versions of the SYSTRAN system. The tapes are seven channel, odd parity, 800 bpi density and EBCDIC characters. The records are 71 character print lines and are blocked 10 records per block.

Output from the SYSDIF program is a listing of the sentences that differ along with a count of the total number of sentences and a count of the number of sentences that differ.

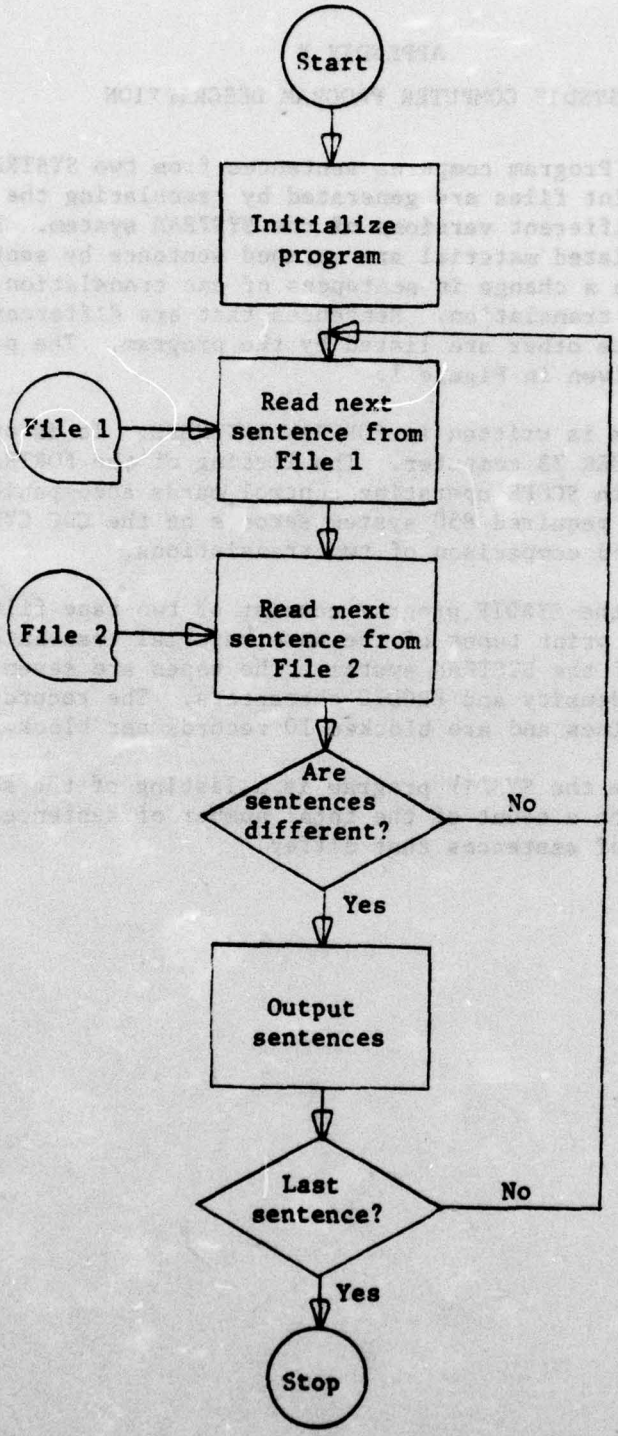

Figure 1. SYSDIF Program Processing Flow

APPENDIX C

AN EXPERIMENTAL MEASURE OF THE COST EFFECTIVENESS OF EXPANDING SYSTRAN SUBJECT AREA DICTIONARIES

The purpose of the machine translation research project being conducted under RADC Contract F30602-75-C-0078 is an evaluation of the SYSTRAN Russian-to-English machine translation system based on the degree of improvement achievable in translation quality of raw system output. This possible degree of improvement in translation quality was achieved through the analysis and upgrading of one of the most basic, but most accessible, components of the machine translation system, the lexical component. The lexical component was analyzed through a concentrated editing effort of a large, representative sample of raw system output. All emendations that fell within the scope of the lexical updating capabilities of the system were introduced into the appropriate lexica of an experimental test system. The lexical analysis and editing were done by bilingual editors, native speakers of Russian, who are also experts in the technical areas covered by the test sample, metals working and mechanical engineering. Further, although the SYSTRAN machine translation system is able to translate technical literature from a number of technical areas, the test samples for this research were drawn from those areas considered to contain the most problems, in the sense of the difficulty and complexity of the subject matter contained in the source language literature, due to the complexity of these technical fields and their highly jargonized nature.

SYSTRAN is not a static machine translation system. Continuous analysis of raw translations produced by the system, either specifically for periodic systems evaluation or indirectly through various post-editing processes, leads to a corresponding continuous updating of the system, mainly in the lexical component. This process, although not of the nature of a concentrated effort like the one of the current machine translation research, does result in a gradual improvement of the lexical component and consequently to an improvement in raw systems output.

Current research involved the production of raw translations of the test samples from two different experimental SYSTRAN systems. The first of these was a system containing an updated dictionary consisting of simple lexical items. The second system contained this dictionary as well as an updated dictionary of semantic expressions. The updating of both dictionaries was based on the corrections to the original text during the editing process. Both of these test translations produced by the experimental machine translation systems were compared with the original output, and analysis of this material has indicated an appreciable degree of improvement in raw output. A major question that now needs to be considered is the following: Since a continual updating of SYSTRAN is always in progress, is such a concentrated updating effort as the one undertaken

during this research cost-effective in terms of system integration within the lexical component, and in terms of improvement of raw system output?

The cost effectiveness of this updating effort as compared against a continuous updating of the system can be considered from at least three different aspects. The first of these is a time and cost analysis of the actual updating of the lexical component of the system, and its direct effect on system output. The second is a study of the effects of the improved lexicon on processes associated with raw system output, namely, the post-editing processes. The third of these aspects is an analysis of the effectiveness of the lexical updating in terms of its applicability, not only to the original test samples, but also to related material, i.e., the cost effectiveness of localized corrections applied on a global scale.

The basis for a cost-effectiveness study on the degree of improvement of raw system output, based only on the analysis of the lexical component of the system, is the test sample of approximately 100,000 words of source language material. This amount of technical literature resulted in 1,316 pages of computer output for the corresponding English translation. Careful editing of this material resulted in approximately 6,400 corrections to the English text, based on a study of the corresponding Russian text. This number of errors reduces to 4.91 errors per computer page printout. Since such a page contains approximately 150 to 180 words, these figures can be refined further to one correction per 30 to 36 words. These figures give a reasonable estimate of the completeness of the lexicon in question. Hence, the question that arises is how the additional 6,400 corrections affected the quality of translation. Careful analysis of the 6,400 corrections eliminated approximately 2,100 as being beyond the scope of the lexical updating. The remainder, roughly 4,300 corrections, were introduced into the system, either into the stem dictionary or into the semantic expression dictionary. The 100,000-word test sample was retranslated, first against the updated stem and old semantic expression dictionary (Phase II), then against both updated dictionaries (Phase III), and the original translation was then compared with each of the new translations. Those sentences that differed in the translations being compared were extracted from the translations and resubmitted for evaluation. The results of these evaluations, in terms of sentences affected, are discussed below.

Of the approximately 7,000 sentences contained in the test sample and the updated Phase II translation, 63 percent were affected by the changes in the lexicon, either directly through the corrections made during the editing process, or through carry-over effects. Of the number of changed sentences, 80 percent were judged to be better than the original. These figures combine to an overall improvement of 50 percent, i.e., fully half of the sentences in the test sample were improved by the lexical updating of the stem dictionary.

Since a large number of the lexical corrections were incorporated into the semantic expressions dictionary, it is essential to consider also the degree of improvement in translation quality when the source material was translated against both updated dictionaries. In this case the number of affected sentences, when the original text was compared with the Phase III translation, increased to 67 percent. Of this number of changed sentences, 86 percent were judged to be better than the original sentences, resulting in an overall improvement of 56 percent of the number of sentences in the test sample. In both cases, the evaluation of the updated sentences was done by the bilingual subject-matter experts who had edited the original text.

TABLE I

INITIAL TRANSLATION VERSUS PHASE II AND PHASE III

	Initial Translation vs Phase II	Initial Translation vs Phase III
Sentences Affected	63%	67%
Sentences Improved	80%	86%
Overall Improvement	50%	56%

It must be noted that the 6 percent increase in improvement due, superficially, to the semantic expressions dictionary is somewhat misleading. The semantic expressions dictionary affected sentences that had been previously improved by the stem dictionary, as well as sentences that had not been previously affected. (When test runs were made on Phase II and Phase III translations of the original test material, 42 percent of the Phase II sentences were affected by Phase III changes.) Further, two important parameters of a natural language translation must be considered in this context, translation completeness and translation accuracy. Although these parameters are difficult to gauge individually, it can be stated in general terms that stem dictionaries contribute more to completeness and general accuracy, whereas a semantics expressions dictionary contributes more to specific, contextual accuracy. The importance of the increase in improvement caused by this second dictionary cannot therefore be underestimated.

The work on the system dictionaries demonstrably led to a 50 percent improvement in translation quality. For the purpose of a cost-effectiveness study, this work can be divided into four major stages:

the editing of the original test sample, the analysis of the emendations for insertion into the system dictionaries, the actual coding of the updates and the subsequent system update, and finally the cost factor of the computer processing, together with the associated operations, e.g., key punching. The first of these stages was done by non-Air Force personnel. The other three were done by the Technical Translation Division (NIT), Foreign Technology Division, Wright-Patterson Air Force Base, Air Force Systems Command.

A time and cost analysis, based strictly on time and cost figures or estimates, is never very accurate when applied to a highly subjective field, such as natural language processing. There are always a number of subjective factors to be considered in conjunction with the objective data. Although the following cost-effectiveness study for the first part of the overall analysis will be based on objective data, some of these subjective factors will be listed also, to be taken into consideration for the overall analysis.

The subject-matter experts doing the machine translation test sample evaluation did their evaluation on the basis of their knowledge of the Russian language, not restricted by the design parameters of a machine translation system. Also, the design of the research effort demanded a careful, thorough, and accurate editing process. Hence a number of corrections were generated that were either unsuitable or impossible for lexical updating. Similarly, because of the unfamiliarity of these experts with the SYSTRAN updating system their rate of editing was slower than the rate of a SYSTRAN staff analyst.* Finally it must be kept in mind that some of the following figures are only best possible estimates, due to the breakdown of the updating process into several specific stages, caused by the different groups of researchers involved in the various stages.

TABLE II

TIME AND COST

	Hours	Cost
Text Editing	999	$13,986
Analysis of Update Cards	360	$ 4,100
Coding and Update	1,320	$10,570
Materials		$ 200
Computer Processing (including key punching)		$ 2,000
Total	2,679	$30,856

*Average rate for subject-matter experts = 6 changes per hour.
Average rate for SYSTRAN staff analysts = 8 to 12 changes per hour.

A reasonably direct measure of cost effectiveness for a machine translation study is the degree of improvement found in translations of related technical areas caused by the updating of dictionaries. That is, any dictionary work can only be cost-effective if the effects of such an updating extend beyond the purely localized material, the material whose editing caused the initial updating. Hence, for the purpose of a carry-over effects study, a test sample of approximately the same size as the original test sample, and drawn from the same technical areas of metals working and mechanical engineering, was translated using the original and the two experimental SYSTRAN systems. These translations were compared against each other, using the same procedures as for the original translations. The evaluation of these sentences that changed in the translations being compared was made by a set of evaluators different from the initial editors. In general terms, since it is practically impossible to match natural language test samples exactly, the expected result was a decrease in the number of sentences judged better in the Phase II and Phase III translations. Nevertheless, the findings of the carry-over effects study showed very positive results. The following table parallels the one given for the original study and is based on the 6,400 sentences of the second text sample.

TABLE III

CARRY-OVER EFFECTS STUDY

	Initial Translation vs Phase II	Initial Translation vs Phase III
Sentences Affected	57%	67%
Sentences Improved	70%	61%
Overall Improvement	40%	41%

There is obviously still a considerable improvement in translation quality. The decrease in improvement in going from Phase II to Phase III translations may have been caused by one of the evaluators whose scores differed markedly from the others. Percentages for the original versus Phase III sentences, excluding the results of this one evaluator, were 71 percent. These results can be combined for the following cost-effectiveness study.

TABLE IV

INITIAL STUDY
(7,000 SENTENCES)

	Original Text vs Phase II	Original Text vs Phase III	Increase Phase III/Phase II
Sentences Affected	63%	67%	6%
Sentences Improved	80%	86%	7%
Overall Improvement	50%	56%	12%

TABLE V

CARRY-OVER EFFECTS STUDY
(6,400 SENTENCES)

	Test Text vs Phase II	Test Text vs Phase III	Increase Phase III/Phase II
Sentences Affected	57%	67%	17%
Sentences Improved	70%	61%	-13%
Overall Improvement	40%	41%	2%

The total number of changes made by the subject-matter editors was 6,400, of which 4,300 were usable for lexical updating purposes. The cost effectiveness of this work will be discussed in terms of the effects these changes had on textual units, the 7,000 sentences of the sample texts.

Total Editing Cost	:	$13,986
Total Updating Cost	:	$16,870
Total Cost	:	$30,856
Total Editing Cost per Sentence	:	$2.00
Total Updating Cost per Sentence	:	$2.41
Total Cost per Sentence	:	$4.41

Since 50 percent of the test sentences were actually improved by the lexical updating procedure, the cost for improving a sentence is actually $8.82. This cost figure, however, does not reflect the amortization of the carry-over effects study. Of the 6,400 sentences in that study, 40 percent were improved. Hence, through no further systems work, a total of 6,060 sentences were improved, considering both the 7,000 original sentences and the 6,400 sentences of the carry-over effects study. The cost per improvement is therefore immediately reduced to $5.09. This process can be extended to subsequent translations with a corresponding decrease in the cost per improvement. There is a rate of diminishing return, although this rate cannot be determined from present information.

The final area of possible analysis for a cost-effectiveness study of the SYSTRAN system is one for which no concrete data exist. It is mentioned only because the lexicographic work done in the current research could have a positive effect on this area. This theoretical impact can be discussed in two parts.

The first of these parts is concerned with the number of emendations actually made by the text evaluators. Of the approximately 6,400 corrections, 4,300 were used in the lexical updating process. Twenty-one hundred were rejected as unsuitable for lexicographic work. However, this set of corrections could be separated into five large groups: typographical errors, corrections requiring lexical routine updating, corrections requiring obvious software modifications, corrections that could, with extended effort, be incorporated into the system, and corrections impossible to implement.* The amount of data gathered for one SYSTRAN lexicon would enable SYSTRAN analysts to pinpoint, and hence to improve, problem areas in the system other than the simple lexical work of this project.

The second of these parts involves the 50 percent improvement in raw output acceptability and the possible effect of this improvement on the post-editing processes. An improvement of 50 percent in raw output acceptability should, at least theoretically, reduce in some degree both the need and the number of requests for post-editing processes.

It is possible to draw a number of conclusions, based on the two factors of translation improvement and cost effectiveness developed in this project.

(1) Translation quality can be improved through lexical updating of a machine translation system.

*Corrections impossible to implement are those that would require changes in the design parameters of the system. An example is a correction that requires analysis of two sentences simultaneously, since SYSTRAN translates on a sentence-by-sentence basis.

(2) Lexical updating produces strong carry-over effects from the test sample to related texts. The rate of improvement appears to be a diminishing one, although the actual rate is not identifiable at present.

(3) Cost effectiveness of lexical updating is due to the carry-over effects of lexical changes with the resulting amortization possibilities of the cost of the initial changes over a period of time.

Based on the same criteria of translation improvement and cost effectiveness in addition to past experience of the Technical Translation Division of the Foreign Technology Division with SYSTRAN, a number of recommendations can also be made.

(1) The same type of study, conducted for the areas of metals working and mechanical engineering, can be conducted for other technical areas with the same type of difficulties, e.g., biology/ medicine, with the expectation of the same degree of success.

(2) Further research of the same type in one of these technical areas could determine the rate of diminishing return for lexical improvement.

(3) The same type of study, conducted in a technical area like physics, which is considered excellent in terms of machine translation work, could determine the upper bound of improvement achievable through lexical updating.

METRIC SYSTEM

BASE UNITS:

Quantity	Unit	SI Symbol	Formula
length	metre	m	...
mass	kilogram	kg	...
time	second	s	...
electric current	ampere	A	...
thermodynamic temperature	kelvin	K	...
amount of substance	mole	mol	...
luminous intensity	candela	cd	...

SUPPLEMENTARY UNITS:

plane angle	radian	rad	...
solid angle	steradian	sr	...

DERIVED UNITS:

Acceleration	metre per second squared	...	m/s
activity (of a radioactive source)	disintegration per second	...	(disintegration)/s
angular acceleration	radian per second squared	...	rad/s
angular velocity	radian per second	...	rad/s
area	square metre	...	m
density	kilogram per cubic metre	...	kg/m
electric capacitance	farad	F	A-s/V
electrical conductance	siemens	S	A/V
electric field strength	volt per metre	...	V/m
electric inductance	henry	H	V-s/A
electric potential difference	volt	V	W/A
electric resistance	ohm		V/A
electromotive force	volt	V	W/A
energy	joule	J	N-m
entropy	joule per kelvin	...	J/K
force	newton	N	kg-m/s
frequency	hertz	Hz	(cycle)/s
illuminance	lux	lx	lm/m
luminance	candela per square metre	...	cd/m
luminous flux	lumen	lm	cd-sr
magnetic field strength	ampere per metre	...	A/m
magnetic flux	weber	Wb	V-s
magnetic flux density	tesla	T	Wb/m
magnetomotive force	ampere	A	...
power	watt	W	J/s
pressure	pascal	Pa	N/m
quantity of electricity	coulomb	C	A-s
quantity of heat	joule	J	N-m
radiant intensity	watt per steradian	...	W/sr
specific heat	joule per kilogram-kelvin	...	J/kg-K
stress	pascal	Pa	N/m
thermal conductivity	watt per metre-kelvin	...	W/m-K
velocity	metre per second	...	m/s
viscosity, dynamic	pascal-second	...	Pa-s
viscosity, kinematic	square metre per second	...	m/s
voltage	volt	V	W/A
volume	cubic metre	...	m
wavenumber	reciprocal metre	...	(wave)/m
work	joule	J	N-m

SI PREFIXES:

Multiplication Factors	Prefix	SI Symbol
1 000 000 000 000 = 10^{12}	tera	T
1 000 000 000 = 10^{9}	giga	G
1 000 000 = 10^{6}	mega	M
1 000 = 10^{3}	kilo	k
100 = 10^{2}	hecto*	h
10 = 10^{1}	deka*	da
0.1 = 10^{-1}	deci*	d
0.01 = 10^{-2}	centi*	c
0.001 = 10^{-3}	milli	m
0.000 001 = 10^{-6}	micro	μ
0.000 000 001 = 10^{-9}	nano	n
0.000 000 000 001 = 10^{-12}	pico	p
0.000 000 000 000 001 = 10^{-15}	femto	f
0.000 000 000 000 000 001 = 10^{-18}	atto	a

* To be avoided where possible.

MISSION
of
Rome Air Development Center

RADC plans and conducts research, exploratory and advanced development programs in command, control, and communications (C^3) activities, and in the C^3 areas of information sciences and intelligence. The principal technical mission areas are communications, electromagnetic guidance and control, surveillance of ground and aerospace objects, intelligence data collection and handling, information system technology, ionospheric propagation, solid state sciences, microwave physics and electronic reliability, maintainability and compatibility.